Master of Science Thesis
2026

An objective assessment of the differences between synthetic and digital mammography using AI

Emma Kristoffersson

Supervisors

Anna Bjerkén, Magnus Dustler and Anders Tingberg

Medical Radiation Physics, Lund
Medical Physics Programme
Faculty of Science
Lund University, Sweden
www.msf.lu.se

Populärvetenskaplig sammanfattning

När kvinnor genomgår en mammografiundersökning så tas olika typer av röntgenbilder för att kunna upptäcka bröstcancer. Regelbundna mammografiundersökningar är viktiga eftersom tidig upptäckt kraftigt minskar dödligheten från bröstcancer. Nya tekniker utvecklas konstant för att förbättra hur väl man kan undersöka bröstet utan att göra en invasiv undersökning, samtidigt som inte vill öka stråldosen. En typ av sådan teknik är den syntetiska mammografibilden (SM) som skapas utifrån en 3D-undersökning. Denna 3D-undersökning kallas digital brösttomosyntes (DBT). En SM-bild produceras i stället för att det skall tas en ny vanlig 2D-bild utöver undersökningen.

I detta arbete så undersöker jag hur väl dessa SM-bilder fungerar när de analyseras av ett AI-system, och hur de presterar jämfört med de traditionella 2D-bilderna (DM) som används i dagens mammografiundersökning. Eftersom SM-bilderna inte har exakt samma format som DM, behövde de först förberedas så att AI-systemet kunde tolka dem likvärdigt. När detta var gjort fick AI-systemet analysera båda bildtyperna, och resultaten jämfördes för att se om SM kan fungera som ett tillförlitligt alternativ till DM.

Analysen visar hur väl detta AI-system kan arbeta med syntetiska bilder, men att SM tenderar att vara mer försiktig i sina bedömningar. I denna studie innebär det att fler kvinnor klassades som misstänkta för bröstcancer trots att de inte hade någon konstaterad sjukdom. Det betyder inte att SM är en dålig metod, men det visar att tekniken behöver utvärderas noggrant innan den kan ersätta DM i klinisk rutin. En alltför försiktig metod kan leda till fler återkallade, vilket kan skapa oro för patienterna och belasta vården. Samtidigt har SM stora fördelar. Eftersom bilden skapas från DBT-undersökningen får radiologerna både 2D- och 3D-information från samma undersökningstillfälle, utan att stråldosen ökar. Detta gör tekniken attraktiv för framtidens mammografi, där målet är att kombinera låg stråldos med hög detaljrikedom.

För att kunna dra säkra slutsatser behövs dock större studier. Den här studien visar tydligt att SM har potential, men också att det finns utmaningar som måste hanteras innan tekniken kan användas fullt ut. Framför allt behövs det mer forskning om hur AI-systemet ska tränas och anpassas för att tolka SM-bilder på ett sätt som motsvarar eller överträffar dagens DM-baserade metoder. Med större datamängder och fler deltagare kan man få mer heltäckande bild av hur SM fungerar i praktiken.

Sammanfattningsvis så visar projektet att SM är en lovande teknik som kan bidra till mer skonsamma och effektiva mammografiundersökningar i framtiden. Men för att tekniken ska kunna införas brett krävs fortsatt forskning, noggranna jämförelser och större studier.

Abstract

Background and purpose: Breast cancer is the most common type of cancer among women in Sweden. The mortality can be significantly reduced if the cancer is detected at an early stage. Screening programs in Sweden are offered for all women aged 40-74 years, using conventional digital mammography (DM). Digital breast tomosynthesis (DBT) is superior to DM in terms of diagnostic accuracy, but has a higher workload than DM. From the DBT image volume, a synthetic mammography (SM) image is created.

The SM image is generated from the DBT to provide a 2D-projection image for radiologists, thereby eliminating the need for a separate DM acquisition and avoiding additional radiation dose. A potential screening strategy would be to image all women with DBT and let an AI system triage the women depending on their breast cancer risk to be investigated by SM (low-risk women, which is a majority of the women in a screening population), or their DBT images (high-risk women). The SM images could potentially be used as a substitute for DM, to achieve a suitable combination of SM for low-risk women, and DBT for high-risk women. This study aims to investigate whether SM created from DBT image volumes is equivalent to DM regarding diagnostic accuracy, by using AI cancer detection software.

Material and Methods: One hundred women recalled from screening were invited to participate in this preparatory study. The women who accepted to participate underwent imaging using two different modalities, DBT and DM. SM was reconstructed from the DBT volumes. The imaging is performed at Unilabs, Malmö, Sweden, using a clinically used mammography screening device (B.brilliant, Siemens Healthineers, Forchheim, Germany). An AI system, Transpara (Screenpoint Medical, Nijmegen, NL) was used to analyze the images. The SM images were preprocessed to match the format required by the AI system, enabling them to be analyzed in the same way as the DM images. The AI-system provided a risk score that represents the level of cancer suspicion. The AI risk scores of the SM and DM with the malignant cases were compared using a statistical test to test the null hypothesis that there is no significant difference between the modalities. Further analysis of all 100 cases was done using ROC analysis to evaluate diagnostic accuracy.

Results: Among the 36 malignant cases, the SM yielded consistently higher risk score than DM. A Wilcoxon signed-rank test confirmed a statistically significant paired difference between modalities ($p < 0.05$). From the ROC analysis of the 100 cases, the resulting area under the curve, AUC, for DM was 0.843 compared to the AUC for SM which was 0.753. Based on DeLong's test the difference in AUC indicated that there was no significant difference between the two modalities ($p=0.071$).

Conclusion: These results indicate that SM provides AI risk scores comparable to, or higher, than DM for malignant cases. The ROC curve for SM showed a tendency toward higher false-positive rates, suggesting increased sensitivity. If further studies confirm that the diagnostic accuracy between the modalities show no significant difference, then these findings support the suitability of SM as a potential substitute for DM images in future screening programs.

Abbreviations

AI – Artificial intelligence

AUC – Area under the ROC curve

CC – Craniocaudal

DBT – Digital breast tomography

DM – Digital mammography

FP – False positive

MLO - Mediolateral oblique

ROC – Receiver operating characteristic

SM – Synthetic mammography

Table of Contents

1. Introduction

Breast cancer is the most common type of cancer for women in Sweden and is the cause of around 1,400 deaths annually (1). The mortality can be significantly reduced if the breast cancer is detected at an early stages (2). The screening program in Sweden is offered for all woman aged 40-74 years, where the women undergo x-ray imaging including two views of each breast using conventional digital mammography (DM). Early-stage breast cancer is more difficult to detect for high-risk women, particularly those with dense breast tissues, since the density reduces the sensitivity of DM (3).

Studies from different research groups have shown that digital breast tomosynthesis (DBT) is superior to DM in terms of diagnostic accuracy (4, 5). From DBT, SM images are reconstructed from the image volumes. DBT is especially better for women with dense breasts, by increasing sensitivity and reducing masking effects of the normal breast tissue (6). However, the radiologist workload is higher with DBT compared to DM (7), which has hampered the introduction of DBT in screening in many European countries.

One feasible strategy would be to examine a majority of women invited to screening with DM (low workload), whereas the small proportion of women with a higher risk for having breast cancer would be examined with DBT (higher workload). This approach would require that you have to know before the imaging whether the woman has high or low risk of having breast cancer.

Another strategy would be to image all women with DBT and let an AI system triage the women depending on their breast cancer risk to be investigated by synthetic mammography (SM) (low-risk women which is a majority of the women in a screening population), or their DBT images (high-risk women). The SM images could potentially be used as a substitute for DM, to achieve a suitable combination of SM for low-risk women, and DBT for high-risk women.

An ongoing screening project, MAST, aims to evaluate the reading time, sensitivity and specificity of a new cancer screening strategy. The new screening strategy includes DBT enhanced with AI, with a first goal to use human reader studies in a clinical setting to establish the equivalency of SM and DM. Previous studies have reported that SM are similar to DM images (8-10) , but if they are exactly equivalent in diagnostic and screening settings is unknown.

1.1 Aim

The aim of this study is to examine SM and DM images of the same breast and investigate if there are any substantial output differences using AI cancer detection software, as a complement to human reader studies.

2. Theory

2.1 Imaging modalities used in screening

Breast cancer screening relies on X-ray based modalities to visualize subtle differences in soft-tissue attenuation such that detection of malignancies at an early stage can be identified (11). Modalities used in mammography operate within the low-energy X-ray range to maximize the photoelectric effect, which enhances contrast between the adipose and glandular tissue. In order to distinguish findings (e.g., calcifications) in soft tissue, having a very high resolution is crucial (~100μm). The choice of target/filter combinations (e.g., Mo/Mo, Mo/Rh, Rh/Rh, W/Rh) is optimized to shape the X-ray spectrum in order to achieve high contrast at acceptable dose levels (12).

The gold-standard modality for breast cancer screening in Sweden is digital mammography (DM), which replaced screen-film mammography due to its superior contrast resolution, lower radiation dose, and digital image processing capabilities (11). Lately, digital breast tomosynthesis (DBT) has emerged as an advanced modality that reduces tissue superposition and improves lesion detection. To avoid the increased radiation dose associated with acquiring both DM and DBT, synthetic mammography (SM) was introduced as a reconstructed 2D image derived from the DBT volume (13). Together, DM, DBT and SM form the contemporary technical framework for population-based screening, balancing diagnostic performance, radiation exposure, and radiologist workload (5, 14).

2.1.1 Digital mammography

Digital mammography is a 2D projection technique in which the entire breast is irradiated in a single exposure. The physics of DM is closely tied to the technology of the detector. Direct conversion detectors, typically based on amorphous selenium (aSe) photoconductor, convert X-ray photons directly into charge, providing high spatial resolution (15) that is particularly beneficial for detecting suspicious findings in mammography. Indirect detectors, using cesium iodide (CsI) scintillators coupled to photodiodes, offer higher yield (16) but, in comparison a slightly lower spatial resolution due to light scattering within the scintillator.

Image quality in DM is governed by the system's modulation transfer function and noise power spectrum, which together determine the ability to resolve fine structures against anatomical noise. Automatic exposure control systems adjust tube current and voltage to maintain consistent image quality across varying breast thicknesses. Despite these optimizations, the fundamental limitation of DM is the overlapping structures within the breast that can obscure lesions or mimic suspicious findings (5, 17). Although DM demonstrates good sensitivity and specificity in large-scale screening programs, there are limitations (e.g., microcalcifications, dense breasts and tissue overlap) (10, 18). In the Malmö Breast Tomosynthesis Screening Trial (MBTST), DBT consistently outperforms compared with DBT, particularly for architectural distortions and small masses (5, 17). These findings highlight the need for more advanced modalities in dense breast populations.

2.1.2 Digital breast tomosynthesis

Digital breast tomosynthesis (DBT) acquires multiple low-dose projections over an angular range, 15-50 degrees, and reconstructs thin slices through the breast. This reduces the masking effect of overlapping tissue and enhances the detection of subtle lesions (19).

These projections are reconstructed into thin slices, usually 0.5–1.0 mm thick, which reduces the masking effect of overlapping tissue and improves visualisation of distortions and small masses (19). The physics of DBT involves a trade-off of angular sampling, projection number and radiation dose. A greater amount of projections improve the depth resolution but increase the acquisition time and dose, while fewer projections reduce the dose at the expense of image quality (20). The reconstruction algorithms are initially filtered back projection and increasingly iterative reconstruction, in order to suppress noise and reduce artifact (21). The slice sensitivity profile determines how well structures are separated in depth, influencing the radiologist's ability to distinguish true lesions from superimposed tissue (19).

Evidence from MBTST demonstrates that DBT yields higher cancer detection rates and lower recall rates than DM, both in initial and subsequent screening rounds (17). Implementation studies from Malmö further show that transitioning to DBT requires adjustments in workflow, training, and technical infrastructure, but provides substantial diagnostic benefits (14). However, DBT generates a significantly larger number of images per examination, increasing reading time and motivating the integration of SM and artificial intelligence to manage workload.

2.1.2.1 Synthetic mammography

Synthetic mammography (SM) is a 2D image generated from the DBT acquisition, by using back-projection and edge enhancement techniques, developed to eliminate the need for a separate DM acquisition (19). This approach reduces radiation dose while preserving the diagnostic advantages of DBT. Comparative studies demonstrate that SM is diagnostically comparable to DM in many scenarios, and that the combination of SM and DBT provides high sensitivity without additional dose (5). More recent work from Malmö shows that AI can support the selection of cases where SM alone is sufficient versus cases requiring DBT review, thereby reducing radiologist workload without compromising diagnostic safety (22).

2.2 Breast cancer screening

About nine out of ten women survive for at least ten years after their breast cancer diagnosis (23). The prognosis is positive, which is partly due to the mortality can be significantly reduced if detected early (2, 24). To enable the detection of early signs of breast cancer, mammography screening is used (24). In mammography, the standard screening method is DM globally (25), although supplemental modalities such as ultrasound, DBT and MRI may be used.

In Sweden, Socialstyrelsen (the Swedish National Board of Health and Welfare) regulates population-based breast cancer screening, nationally in Sweden (24). The screening is offered to all women aged between 40-74 years with an interval of 18-24 months. The statutory regulations and evidence-based guidelines that governs the national breast cancer screening program is Socialstyrelsen (24, 26) and Nationellt vårdprogram för bröstcancerscreening, defines the target populations, quality indicators, and evidence-based recommendations (27).

The screening examination is performed using DM imaging. The imaging procedure consists of acquiring two standard views: the craniocaudal (CC) view and the mediolateral oblique (MLO). The CC view is performed when the mammography unit is at a zero-degree position and the MLO view is performed at an angle, as illustrated in Figure 1.

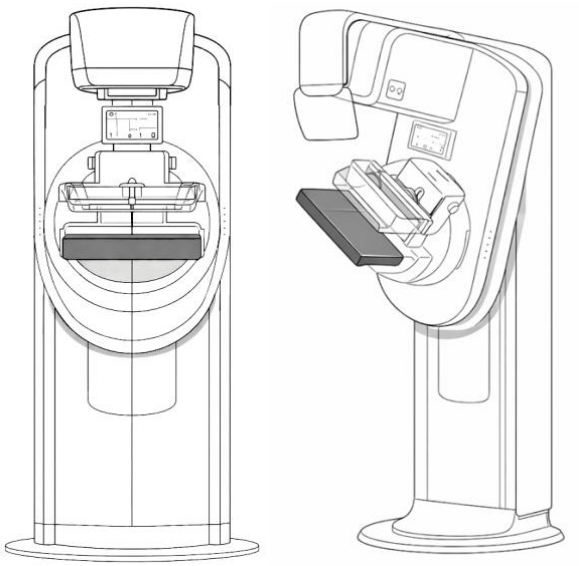

Figure 1. Line art illustration adapted from the manufacturer's manual (Siemens Healthineers B.brilliant (28)). The left image is a CC view (head-to-feet), and the right image is an MLO view (parallel to the pectoralis muscle). The part that is darker illustrates the detector and the compression plate is above.

In the CC view, the low-energy X-ray photons enter the breast superiorly and exit inferiorly. This projection helps determine whether lesions are located medial or lateral to the nipple, and provides information along the anterior-posterior direction of the breast. In the MLO view, the breast is positioned so that the pectoralis muscle is visualized extending parallel down towards the nipple. This view provides further information about the tissue in the axillary region (29).

Before imaging, the breast is positioned between the detector and the compression plate. The compression plate is used to flatten the breast in order to minimize the scattered radiation, reduce motion, and also to obtain higher-quality images (30). This procedure is then repeated so that both breasts are imaged.

2.2.1 MAST

Sweden continuously evaluates the outcomes of screening, including participation, recall rates, and interval cancers, and is currently acknowledging that there is a potential to use artificial intelligence (AI) in screening (31). This is a hot topic for several studies within the field, one of them is Malmö AI-based breast Screening with Tomosynthesis (MAST). This MAST study is ethically approved, and written informed consent is obtained from all participants prior to inclusion.

In a preparatory study for MAST, women recalled from screening are invited to participate. Once in the study, the women are imaged using two different modalities, DBT (as illustrated in Figure 2) and DM. The same views are then taken together with the DBT for each breast. From the DBT image volume, a SM is created.

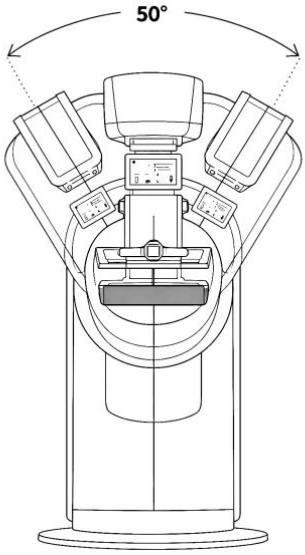

Figure 2. Line art illustration adapted from the manufacturer's manual (Siemens Healthineers B.brilliant (28)) performing a DBT in a CC view, with an angular range of 50 degrees.

The imaging is performed at Unilabs, Malmö, Sweden, using a clinically used mammography screening device (B.brilliant, Siemens Healthineers, Forchheim, Germany).

2.2.1.1 Inclusion

Based on a power calculation for the upcoming human reader study, 300 women will be included in the study. The current pilot project included 100 of these women. Since the patients in this study are recalled from screening due to cancer suspicion, the case set was cancer enriched, including subtle and evident lesions (approximately 40/60 distribution between cancer and benign/normal). The cancer cases were confirmed by pathology reports, and normal cases by follow-up over three to six months). Exclusion criteria for this study included pregnancy, implants, piercings or language barriers.

2.3 Artificial intelligence

Artificial intelligence has been increasingly implemented across a variety of different sectors, including finance, transportation, and healthcare. AI-based systems are used for tasks such as image interpretation, risk prediction, workflow optimization, and decision support within medicine (32). One area that would be very beneficial of implementing AI is breast cancer screening, since there are a great amount of imaging data and the need for consistent interpretation (33).

In mammography screening, AI systems are typically used to analyze the DM, SM, and DBT images and provides a score that represents the suspicion of malignancy risk (33). Research has shown that AI can perform at a level comparable to a radiologist (34), and several studies have demonstrated that AI may reduce workload, act as a second reader, or triage low-risk examinations for single-reader review (35, 36). Although AI is not yet implemented in routine screening in Sweden, national authorities acknowledge its potential role in improving efficiency and diagnostic performance (27).

2.3.1 Screenpoint Transpara

The AI system Transpara (Version 2.1.1, ScreenPoint Medical, Nijmegen, Netherlands), is a commercially available deep-learning based AI systems designed to support breast cancer detection in mammography (37). The system analyses each mammography image using convolution neural networks trained on large datasets of digital mammography examinations, which does not include SM images (34, 38). The information from the four images originating from the screening-procedure (R/L-CC and R/L-MLO) is taken into consideration in order to generate an AI risk prediction.

Transpara uses an AI image-analysis algorithm to evaluate the breast tissue and mark any suspicious findings that could potentially indicate malignancy (38). Each suspicious region is assessed individually and assigned a highest region score ranging from 1-100, were 100 is indicating the greatest malignancy suspicion.

The evaluation also generates an overall exam score, or Transpara score (TS), which is a malignancy score on a scale of 1-10, and represents the estimated likelihood of cancer (38). An individual TS for each of the four views is presented together with a comprehensive evaluation that incorporates all views. The resulting evaluation can also highlight suspicious regions through region-based detection marks that indicate areas contributing to the elevated risk assessment (38). These marks are intended as a guide for the observer (e.g., radiologist).

According to Rodriguez-Ruiz et al. (34), Transpara can operate either as a stand-alone reader or as a decision-supporting tool that can be integrated in a double-reading workflow. In a large multi-reader study involving 101 radiologists, Transpara demonstrated a performance within the range of human readers (34). When Transpara has been used as decision-support, it has been shown to improve the cancer detection and reduce the reading variability (34, 38).

3. Materials and methods

3.1 Data

The images (all views of SM and DM) were analyzed with Transpara and assigned a Transpara score (TS), 1-10, that represents the level of malignancy risk, where 10 represents the highest level of suspicion. The results from the AI-system were used to compare the diagnostic accuracy of the two modalities (DM and SM).

For each woman, the AI-system produces a comprehensive evaluation that incorporates all four views of the examination (two views per breast). This evaluation is performed independently for the DM, SM and DBT modalities. The TS is calibrated so that approximately 10% of cases fall into each score level (1-10). The comprehensive evaluation is assigned a TS, 1-10, that is connected to the highest region score in all four images. If the comprehensive evaluation gives a TS of 10, then it represents a highly suspicious finding which would most likely result in a cancer diagnosis. If the analysis gives a lower TS than 10, then it is classified as benign or normal in these results. The malignant cases were confirmed through pathology reports, and benign/normal cases had a follow-up period of 3-6 months.

Two diffent analysises were performed using ROC analysis, both based on the same set of patients. These were:

- **DM vs. SM** for the malignant and benign/normal cases
- **DBT vs. SM and DM** for the malignant and benign/normal cases

These anlysises were selected to aseess the differences between DM and SM, and for completeness of this pilot study, DBT was also included to provide a full comparison within this dataset.

In addition, we were interested in how SM compared with DM specifically for the malignant cases. Therefore, a separate analysis included only the malignant cases was performed.

3.1.1 Both malignant and benign/normal cases

The analysis consisted of a comparison between SM and DM using ROC analysis curves, for both malignant and benign/normal cases, as illustrated in Figure 4. Additionally, a comparison between the SM, DM and DBT was performed using ROC analysis curves. The research question of this project was a comparison of the AI risk scores (TS) for the two modalities of the same woman in an ROC analysis together with a DeLong's test to assess the significance of differences between the areas under the ROC curves (AUC).

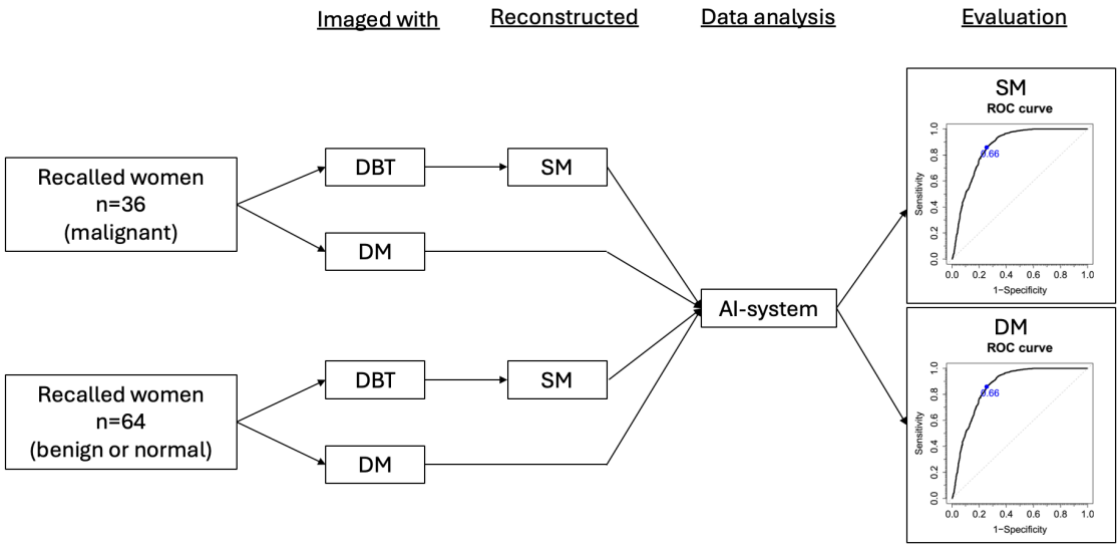

Figure 3. Schematic presentation of the study; a comparison of diagnostic accuracy of DM vs. SM.

3.1.2 Exclusively the malignant cases

The analysis consisted of a comparison of the AI risk score (TS) between the modalities for the same woman and the comparison between the SM and DM was performed using a statistical test, as illustrated in Figure 4.

Before selecting the statistical test, the distribution of the paired differences between DBT and SM/DM, was assessed. These differences were not normally distributed. In addition, the TS is a discrete, ordinal measure that often exhibits skewness, they were not appropriate in this context. As the comparison involve paired observations from the same woman, and the assumptions of normality were not met. For these reasons the Wilcoxon signed-rank test was selected, as it is well suited for skewed or non-Gaussian data such as the TS values.

The cases where the difference in TS differs the most (1 for DM vs. 10 for SM) for the same woman was investigated by consulting a radiologist to get a possible statement on the image quality as evaluated by a trained human observer.

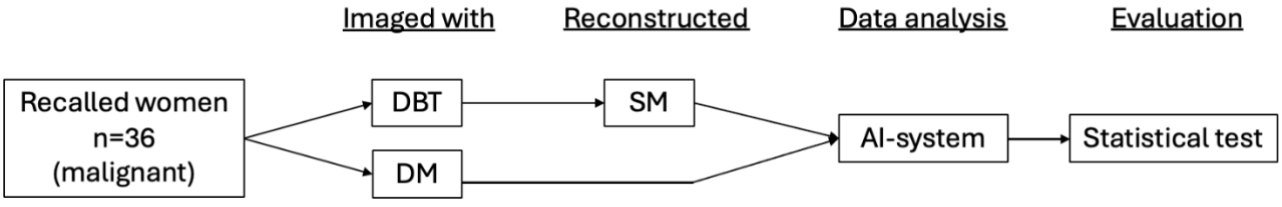

Figure 4. Schematic presentation of the study; a comparison of diagnostic accuracy of DM vs. SM.

3.2 Image processing

The AI-system is designed to read and analyze DBT volumes alongside an accompanying SM or DM image. While DM images can be processed on their own, the AI-system will not accept SM images without the accompanying DBT images. For this study to compare the SM to DM, the AI-system needed to explicitly process only SM. To accomplish this, the SM images were presented in the same format as DM images, by changing the DICOM-header and the image background, as shown in Figure 5. The change of the image backgrounds was necessary as DBT and SM images were cropped, unlike DM which used a set field of view.

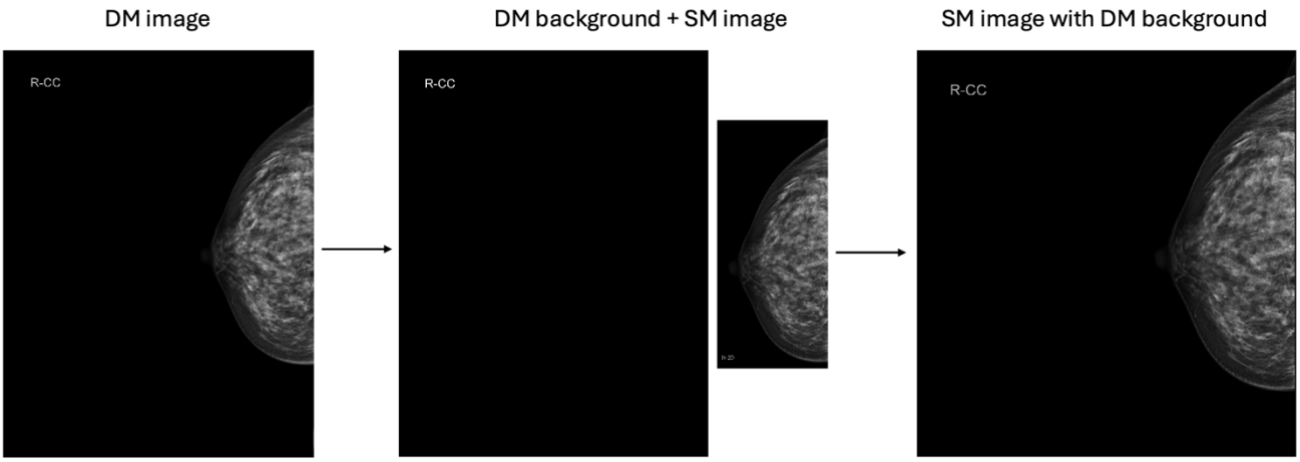

Figure 5. Schematic representation of how the SM image was presented. Segmentation of the DM image was used to remove the object (breast) in order to save the plain background for further use. The corresponding object (breast) in the SM image was removed from the SM background and then inserted into the DM background using similar segmentation. When the image is complete, further DICOM-header changes were performed in order to present the image to the AI-system.

3.2.1 Segmentation

All images were processed using a threshold-based segmentation approach. Each image was first binarized using a simple intensity threshold (BW = image > 0), after which individual objects were identified using connected component labeling, as shown in Figure 6.

The manufacturer appears to apply a preliminary segmentation step, as the background outside of the breast is not uniformly zero. This can be observed in the top-left corner of the CC view in Figure 6. When generating the mask (to segment the breast), a simple threshold was sufficient to separate the breast from the background, indicating that the background contains a low-intensity non-breast pixels rather than true zero values. These residual background structures differ between DM, DBT and SM due to modality specific cropping and preprocessing. To prevent such modality-dependent background patterns from influencing the ROC analysis, the background was standardized across all modality images.

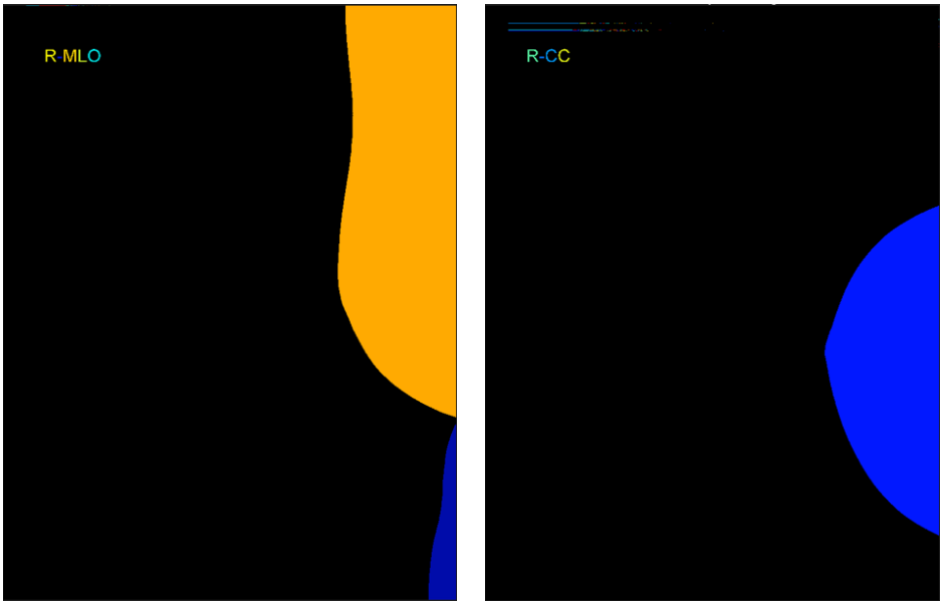

Figure 6. Examples of how the objects were identified for both an MLO (left image) and CC view (right image). The different colors represent a different objects in the segmentation process.

The area for each connected component was computed using regionprops, which measures properties of image regions in MATLAB (R2025b, The MathWorks, Inc., Natick, MA, USA), and the objects were subsequently sorted in descending order of area-size. This allowed the largest component/components (the breast) to be distinguished from smaller unintended objects, as illustrated in Figure 7.

The segmentation was used to remove unwanted objects from the DM (the breast) and to isolate the largest object in the SM image. The bounding box of the SM object was extracted, and its position was adjusted relative to the DM image depending on the desired alignment mode (top, bottom or center/cropped). Pixel values from the SM object were then inserted into the clean DM background to generate the final SM-aligned image.

Since each case represented a unique woman, the entire segmentation and object-handling procedure was performed separately for all four views (R/L-MLO and R/L-CC). All views were visually inspected at each stage to ensure that artifacts, unintended objects, or improperly acquired images were identified before segmentation.

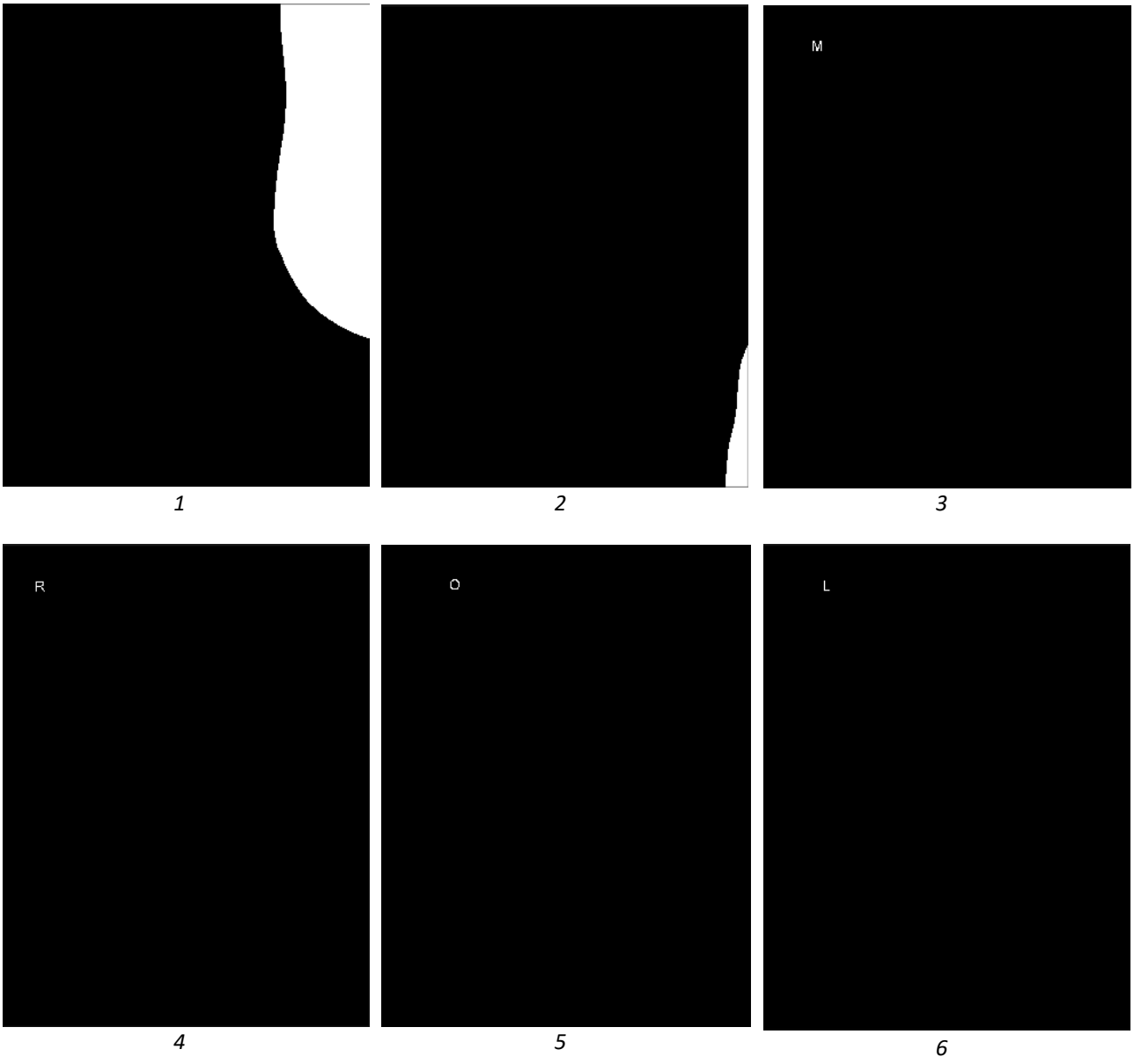

Figure 7. Illustration of the different objects is based on their area of the object within the image. The example shown is of an R-MLO view. The six images are ordered by descending area: the largest object starting with no. 1, the top left and ending with no. 6, in the bottom right which has the smallest area of these elements.

To ensure consistency between DM and SM images, the number of objects removed from the SM image was matched to the number of objects present in the corresponding DM image. Objects were removed based on their area ranking, as illustrated in Figure 7. Finally, all processed images were exported as DICOM files while preserving the original metadata structure.

3.2.1.1 Checkpoints

At each stage before segmentation, all views were visually reviewed to detect any potential problems beforehand. Potential problems included artifacts, unintended objects (e.g., piercings), or issues arising from improperly acquired images that, under clinical conditions, would have required a retake. Such cases would result in differences in breast positioning between the DM and SM images.

In some cases, objects were present in the SM image that did not appear in the corresponding DM image, either due to the examination procedure or the fixed field of view of the DM acquisition. To ensure that the SM image used the same background as the DM image, the number of objects removed from the SM image was adjusted to match the number of objects present in the DM image. The right number of objects was removed by ordering all detected objects by area, as illustrated in Figure 7, and excluding the object that was not in the corresponding DM image into the mask in the segmentation process.

3.3 Placement of SM images

The placement of the SM images is of importance since SM does not always reproduce the same breast geometry as DM. DM is a 2D-projection of the breast, whereas SM is a 2D-projection created from the DBT volume, which is acquired over a range of angles of the breast. Since SM represents a computational projection of a 3D volume dataset rather than a physical projection, the two images may differ in breast outline, tissue distribution, and lesion appearance (19). Clinically, such difference can potentially make DM and SM appear less comparable than expected, and potentially influencing both radiologist interpretation and AI-based assessment.

During the image processing of the synthetic mammography images, a problem occurred since the DM used a set field of view. During the DBT, the imaging swipes over a range of 50° wide-angle DBT with 25 projections (28). Since the DBT swipe is not the same fixed set field of view, there was a smaller field of view for the SM compared to DM. A decision was made that the SM and DM have the same size of field of view and therefore the placement of the SM images in the DM background is modified for the CC-view.

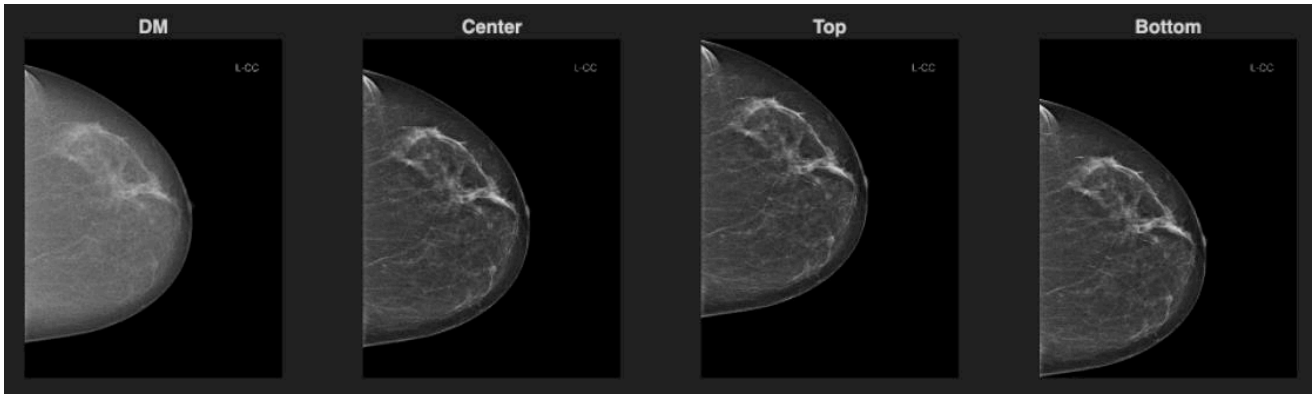

Figure 8. A snapshot of one of the later visual checkpoints. The image on the left is the DM image and the three images to the right of the DM image are SM images.

When using the same method in the MLO-view, a problem occurred. In the MLO-view, the set field of view is more predominant since the other parts of the body (e.g. stomach) is more likely to be imaged in the view and therefore another decision was made, so that if this occurred the image will be cropped so that the SM image is the full field of view of the DBT. This is illustrated in Figure 9. If this was not the case, the full field of view was used (as seen in the CC-view, Figure 8).

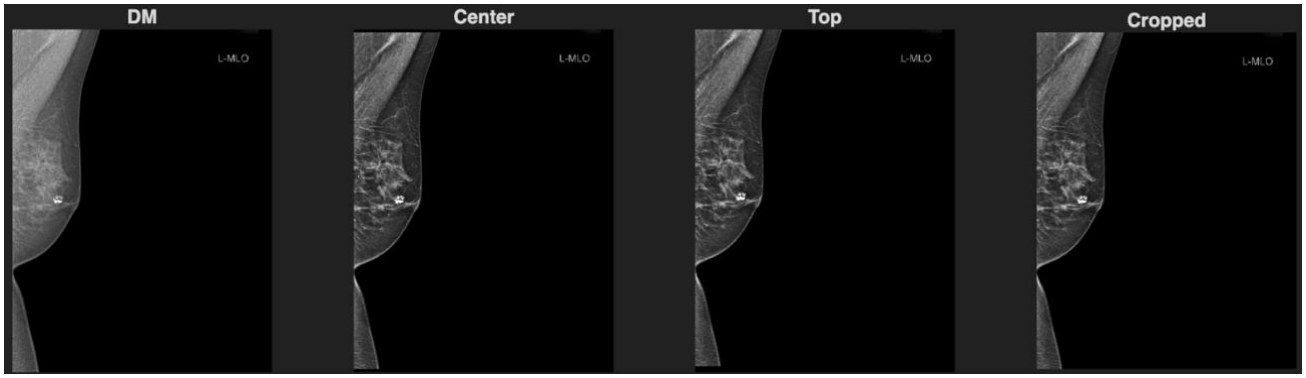

Figure 9. A snapshot of one of the later visual checkpoints. The image on the left is the DM image and the three images to the right of the DM image are SM images.

Statistical tests were performed between the three different placements of the SM image. The corresponding TS for the three placements (top, center and bottom/cropped) for each view represents three repeated measurements of the same underlying image, derived from the same patient and the same projection. The data is therefore paired and not independent. Since TS is an ordinal measure and did not meet the assumptions of normality, a non-parametric approach was required.

To assess whether the placement variants resulted in different TS within each view, Friedman's test was used, which is the non-parametric equivalent of a repeated measures ANOVA and is specifically designed for comparing more than two related samples. To quantify the magnitude of the same-

patient-differences, a Kendall's W test was performed, which is an effect size measure that can be used together with Friedman designs.

When the Friedman test indicated a statistically significant overall difference, a post-hoc pairwise comparison (Wilcoxon signed-rank test) was performed since it is suitable for paired, non-parametric data. The Bonferroni correction was also applied to the Wilcoxon p-values to adjust for multiple pairwise comparisons.

3.4 Analysis of data

All DM and SM images were analyzed by the AI system, and the Transpara scores were recorded. These scores were used in statistical tests for the placement of the SM, malignant cases and the analysis of all included cases was performed by a ROC analysis.

3.4.1 ROC analysis

A Receiving Operating Characteristic (ROC) curve is a commonly used measure of diagnostic performance (39). ROC analysis is based on how well a test or model classifies individuals into positive or negative outcomes. Each case is categorized into one of the four possible outcomes:

- True positives (TP) – a positive case correctly identified
- False negative (FN) – a positive case incorrectly classified as negative
- True negative (TN) – a negative case correctly identified
- False positive (FP) – a negative case incorrectly classified as positive

The ROC analysis is also used to illustrate the trade-off between sensitivity and specificity (39). Sensitivity is the probability of a positive case is correctly identified (40), and is defined as

$$Sensitivity = \frac{TP}{TP + FN}. \quad (2)$$

Sensitivity is plotted on the y-axis of the ROC curve. Specificity is the probability that a negative case is correctly identified (40). For the ROC curve the specificity is used on the x-axis, represented as (1 – specificity) and is also referred to as false-positive rate. The specificity is defined as

$$Specificity = \frac{TN}{TN + FP}. \quad (3)$$

In this project, ROC analysis is performed using MATLAB (2025b). A ROC curve illustrates how the sensitivity and specificity changes as the decision threshold varies (41). In order to objectively compare different diagnostic methods, the area under the ROC-curve (AUC) is used as a summary measure of performance.

3.4.2 Statistical tests

All statistical analysis were performed using MATLAB custom written scripts developed specifically for this study, without dependencies on different toolboxes. The statistical tests that are used in this project include Wilcoxon signed-rank test for the malignant cases, DeLong's test for the AUC analysis and Friedman's test together with a complementary Kendall's W test for the placement of the SM images. All hypothesis testing was performed using a significance level of $\alpha = 0.05$, meaning that p-values below α were considered statistically significant.

4. Results

4.1 Effect of the placement of SM image

For each of the four views, the TS were compared across the three placement variants of the same view (top, center and bottom/cropped). The Friedman's test showed statistically significant difference between placements for all views ($p < 0.05$). However, effect sizes were small (Kendall's $W = 0.077$-0.117), indicating that the magnitude of these differences was limited. Three pairwise post-hoc Wilcoxon signed-rank test were performed. After applying Bonferroni correction, none of the comparisons remained significant ($p > 0.45$).

Table 1. Output from the Friedman, Kendall's W and the Post-hoc Wilcoxon signed-rank test that showed the distribution of the pairwise differences in placement. For the Wilcoxon analysis, the following abbreviations are used: c – center, t – top, and b/c – bottom/cropped

View	L-CC	L-MLO	R-CC	R-MLO
Friedman	p = 0.0004517	p = 2.567e-05	p = 7.909e-06	p = 8.335e-05
Kendall	W = 0.077	W = 0.106	W = 0.117	W = 0.094
Wilcoxon (c vs. t)	p = 0.4524	p = 0.9837	p = 1	p = 0.8817
Wilcoxon (c vs. b/c)	p = 0.7392	p = 1	p = 1	p = 0.8569
Wilcoxon (t vs. b/c)	p = 1	p = 1	p = 1	p = 1

Although the Friedman test detected a statistically significant variation between the three placement variants within each view, the effect sizes were small and none of the pairwise comparisons reached significance. This suggests that the TS are largely stable across center, top and bottom/cropped placements within the same view, and that the observed difference are unlikely to be clinically meaningful. The global variation (views in Table 1) is statistically detectable, but no individual pairwise comparison (the last three lines in Table 1) is large enough to stand out on its own.

4.2 Comparison of SM and DM

From the 100 cases, DM demonstrated a slightly higher discriminative ability than SM. The resulting AUC for DM was 0.843 compared to the resulting AUC for SM which was 0.753. Based on DeLong's test the difference in AUC indicated that there was no significant difference between the two modalities (p = 0.0713).

The ROC curves for both modalities are shown in Figure 10. AUC values derived from the trapezoidal method were slightly lower, as expected for discretized scores, yielding an AUC = 0.836 for DM and AUC = 0.747 for SM.

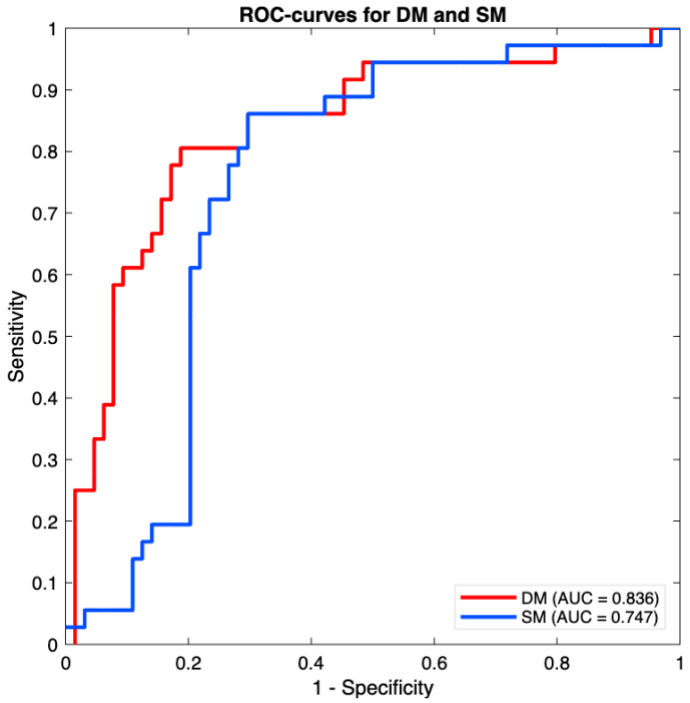

Figure 10. The resulting ROC curves for DM and SM. The AUC was determined by the trapezoidal rule.

4.3 Comparison of SM, DM and DBT

For completeness, Table 2 includes the DM vs SM results previously reported in the preceding section.

Table 2. Output from the DeLong's test showed the corresponding AUC values for DM, SM and DBT together with their respective p-values.

Comparison	AUC (95% CI)	p-value
DBT vs. DM	0.859 (CI: 0.804-0.915) vs. 0.843 (CI: 0.762-0.924)	0.7553
DBT vs. SM	0.859 (CI: 0.804-0.915) vs. 0.753 (CI: 0.671-0.836)	0.0215
DM vs. SM	0.843 (CI: 0.762-0.924) vs. 0.753 (CI: 0.671-0.836)	0.0713

The DeLong's test indicated that DBT and DM performed similarly, with no statistically significant

difference in AUC (0.859 vs. 0.843; p = 0.7553). In contrast, DBT showed a significantly superior discriminative performance compared with SM (0.859 vs. 0.753; p = 0.0215).

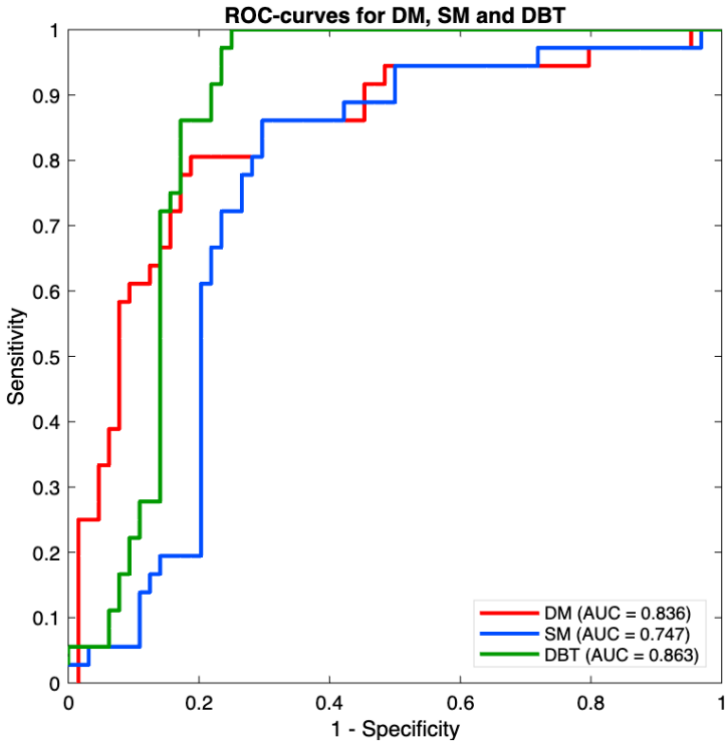

Figure 11. The resulting ROC curves for DM, SM and DBT. The AUC was determined by the trapezoidal rule

4.4 Comparison of SM and DM for malignant cases

Of the 100 included cases, 36 cases had confirmed malignancy. A Wilcoxon signed-rank test showed a statistically significant difference between DM and SM TS ($p = 3.597 \times 10^{-6}$).

The score distributions differed markedly between modalities. The SM frequently assigned the maximum score to the malignant cases (TS = 10), whereas DM scores were spread across the full range (TS 1–10). This pattern is summarized in Table 2, which shows a concentrated distribution for SM towards the highest TS values, and a more heterogeneous distribution for DM.

Table 2. The distribution of resulting TS (1-10) for DM and SM, of the malignant cases.

TS	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
DM	2	3	2	2	5	2	6	2	4	8
SM	0	0	0	1	0	0	1	0	3	31

A paired comparison of the TS on the y-axis in Figure 12, for the same women using DM and SM, Figure 12, further illustrates the systematic upward shift in SM scores. Although the data are paired and expected to be similar, several cases showed substantial discrepancies.

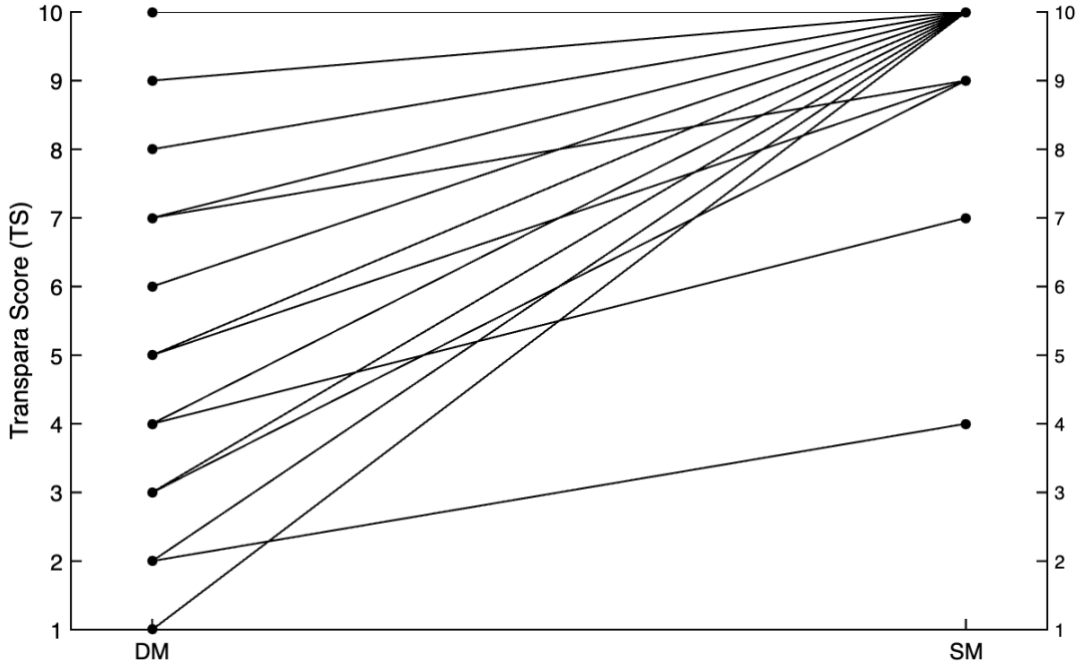

Figure 12. Paired comparison plot of TS for DM vs. SM for the malignant cases.

Notably, two malignant cases demonstrated extreme discordance, with DM assigning TS = 1 while SM assigned TS = 10. The radiologist reviewed these two cases to provide an additional professional perspective on these findings and offer a possible statement to why the AI system produced remarkably different scores. The radiologist had access to the images (the four views of each case) and no other information other than that the cases were confirmed malignancies. A viewing software that allowed standard diagnostic tools such as windowing, magnification, and adjustments of contrast such that a more detailed assessment could be performed.

The radiologist overall statement of these two cases was that the SM images are easier to evaluate due to the higher contrast and was considered the main concrete reason for distinguishing the two modalities. In one of the cases a distortion in the breast could be identified in the MLO-view, as shown in Figure 13, for both modalities and therefore would be classified as a malignancy category four.

In the other case, a suspicious mass (Figure 14), could be identified in the CC-view by the radiologist. In this case, the mass was less clearly delineated than the previous case according to the radiologist and was therefore classified as a malignancy category of three or four.

In the structured report of the SM images, which is provided by the AI system after an evaluation if there are findings, the highest score was of a calcification and not the malignant transformation, which had a lower score. According to the radiologist, the calcification could be identified in the SM image, but not in the DM image. Across all malignant cases, calcifications were marked more frequently in SM than in DM. While DM yielded only 6 identified calcifications, SM revealed 29 which indicates a systematic visibility advantage for the SM.

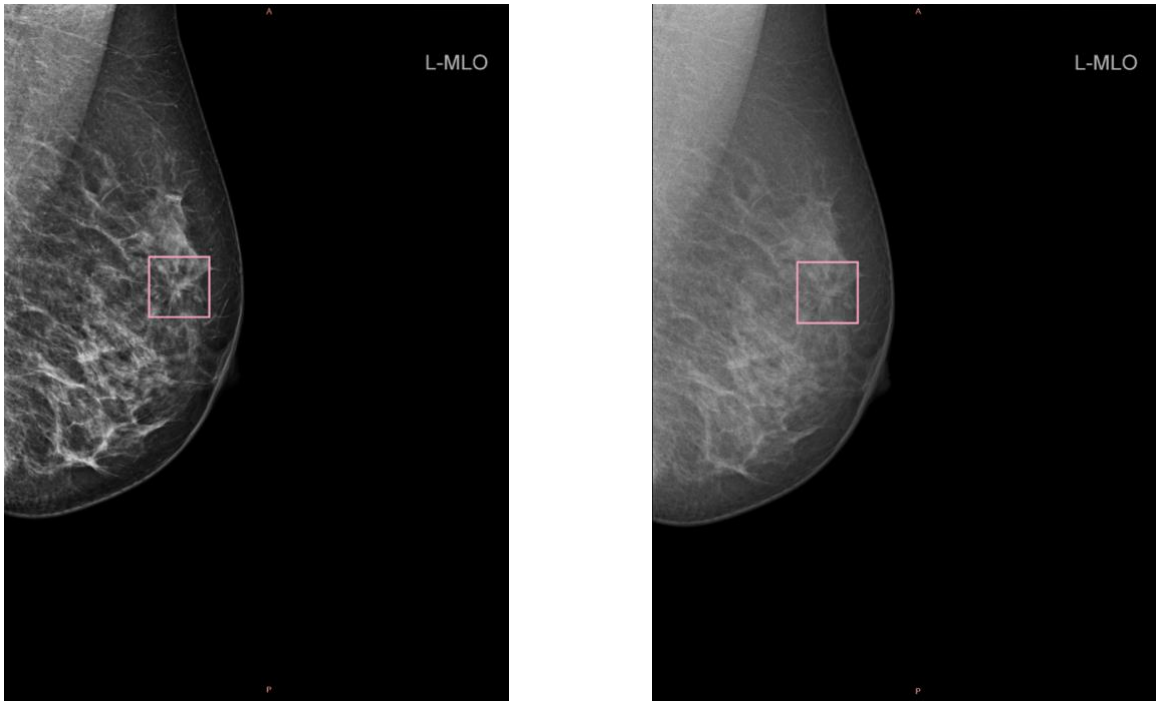

Figure 13. One of the cases, where TS = 1 for DM and TS = 10 for SM (without windowing). The figure to the left is SM and DM is on the right. The finding that the AI-system marked is shown inside the pink, and according to the radiologist it is the same location for the finding In the DM image.

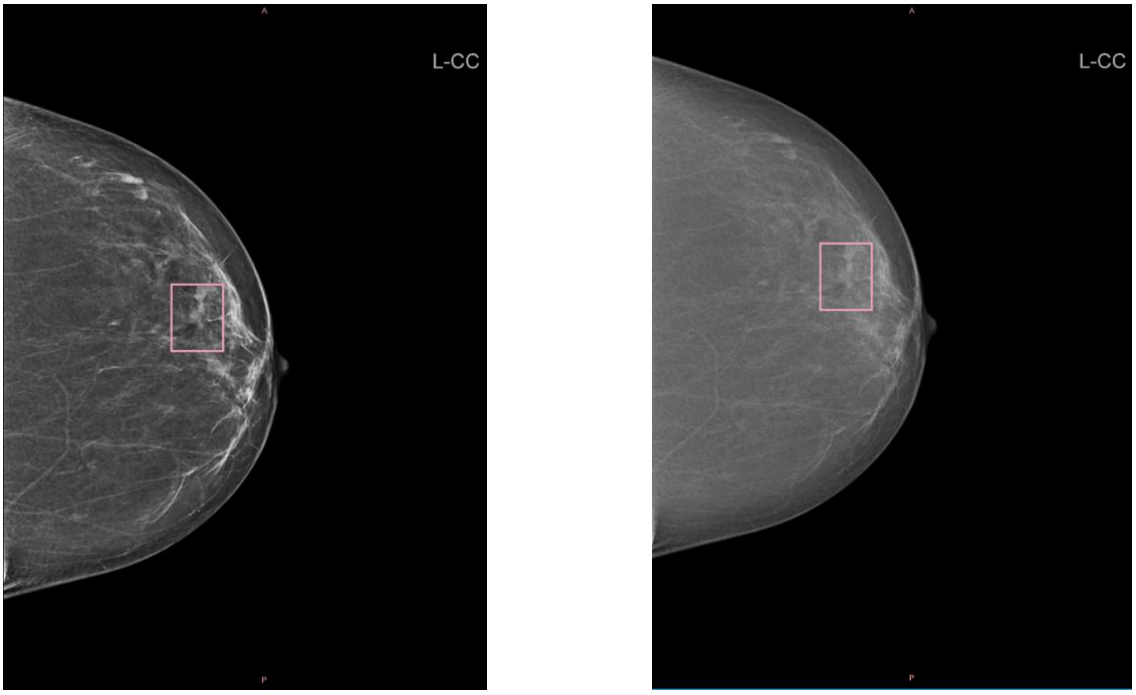

Figure 14. The second case, where TS = 1 for DM and TS = 10 for SM (without windowing). The figure to the left is SM and DM is on the right. The finding that the AI-system marked is shown inside the pink, and according to the radiologist it is the same location for the finding in the DM image.

5. Discussion

The proposed approach provides an alternative and potentially more objective means of assessing the differences between SM and DM images, complementing the traditional evaluations based on human observers. In this preparatory study, the comparison of paired SM and DM images with confirmed malignancies revealed measurable differences in AI-derived output, with SM consistently producing higher TS than DM for the same woman. The statistically significant shift in score distribution indicates that the two modalities are not interpreted equivalently by the AI system, reflecting underlying differences in image characteristics between SM and DM. These findings highlight that modality-specific behavior of AI systems must be considered when integrating SM into AI-supported screening workflows.

5.1 Comparison of SM and DM

The ROC curve for SM demonstrates higher false-positive rate at comparable sensitivity levels, which according to ROC theory reflects reduced specificity (39). In the context of breast cancer screening, an elevated false-positive rate is associated with increased recall rates and a tendency toward overcalling of suspicious findings (27).

Despite the lower AUC for SM, the operating characteristics of SM suggest that it behaves as a more sensitive, false-positive prone modality. This pattern is consistent with a system that is cautious and sets a positive diagnosis of suspicious findings. In a clinical workflow, such behavior is often desirable as a compliment to the reader, where increased sensitivity is prioritized over specificity in a population based breast cancer screening (25).

Although DM and SM are both 2D-mammographic formats, they differ fundamentally in image formation. SM is synthesized from DBT data and therefore exhibits distinct intensity characteristics, noise properties, and structural patterns compared with DM. These differences create a domain shift that can degrade the performance of AI systems trained on DM, unless the models are explicitly adapted for SM. Prior studies have demonstrated that AI algorithms trained on DM do not transfer seamlessly to SM, highlighting the need for modality-specific optimization (42).

This difference in image properties likely contributes to the lower AUC observed for SM and its tendency toward higher false-positive rates. Consequently, the performance pattern seen in this project aligns with the expected behavior of an AI-system encountering a different input than the trained data.

5.2 Comparison of SM, DM and DBT

DBT and DM demonstrated comparable diagnostic performance, with overlapping 95% confidence intervals and no statistical difference in AUC. This suggests that DBT did not provide a measurable improvement over DM within this study. However, DBT showed a clearly superior discriminative ability compared with SM, supported by both a higher AUC and non-overlapping confidence intervals.

5.3 Comparison of SM and DM for malignant cases

This study demonstrates that SM systematically produces higher AI-derived malignancy scores than DM when evaluating the same breast. One can see that even when the SM is giving it a lower score, it is still greater than DM, which is illustrated in Figure 7. The significant difference in the paired analysis indicates that this effect is consistent across these malignant cases.

The results show that SM systematically produces higher AI-derived malignancy scores than DM for the same patient, indicating that the current threshold (TS = 10) is perhaps not interpreted equivalently across modalities. This modality-specific score shift, together with the higher false-positive tendency and lower AUC for SM, suggests that the AI system encounters a domain shift when processing SM. It would therefore be valuable to compare these findings on different AI systems, particularly those trained on SM image data, to determine whether the systematically higher AI-derived scores would behave the same.

Although, this preparatory study is not sufficient to define a new operating threshold, these findings indicate that modality-specific calibration, could potentially improve the clinical accuracy. Further studies with larger, population-based datasets are needed to validate these observations and to determine whether a modified threshold for SM can reduce the false-positives while maintaining sensitivity.

6. Conclusions

Although SM demonstrated lower diagnostic performance than DM, the difference did not reach statistical significance, and the confidence intervals showed substantial overlap. The ROC curve for SM displayed a tendency toward higher false-positive rates, consistent with a more sensitive but less specific behavior.

If future studies confirm that the performance gap between SM and DM is not clinically meaningful, the present findings support the possibility of using SM as a substitute for DM in screening contexts.

Acknowledgements

I am especially thankful to my supervisor; **Anders Tingberg**, for your commitment, and making this project possible, and for your continuous support and encouraging guidance throughout the process. I would also like to extend my gratitude to **Magnus Dustler** and **Anna Bjerkén** for your valuable input, constructive feedback, and for helping me strengthen the scientific quality of this work.

I am deeply grateful to my fellow master's student, **Matilda Bilén** and **Jonas Magnusson**, for walking through this entire process with me. Thank you for every shared thought, every exchanged idea, and for all the stress, humor, and laughter that made the toughest days manageable.

I am thankful to the members of the research group, **LUCI** for welcoming me, providing insightful feedback, and offering supportive and motivating words during my time with you.

I would like to express my sincere gratitude to **Akane Ohashi**, for contributing a radiological perspective on the cases that demonstrated extreme discordance in the malignant findings. A special thanks to **Victor Dahlblom** for introducing me to the AI-system and guiding me through its practical use.

I would also like to thank **Helena Erixon** for making it possible for me to shadow the staff at Unilabs during the whole procedure for the MAST study, and for providing essential information. I am also grateful to the personnel at **Unilabs SUS Malmö** for generously sharing your clinical expertise with me.

Finally, I would like to express my deepest appreciation to my family for your love and support, especially my brother's children, who made my Mondays a little less heavy. And of course, a heartfelt thank you to my friends for your constant encouragement.

This work was supported by Region Skåne (Region Skåne Regionalt forskningsstöd and USVE-utrymme), Allmänna sjukhusets i Malmö Stiftelse för bekämpande av cancer, and Siemens Healthineers.

References

1. Socialstyrelsen. Statistik om bröstcancer. Stockholm: Socialstyrelsen; 2023. Report No.: 2023-10-8807.
2. Organization WH. Breast cancer Geneva, Switzerland: World Health Organization; 2026 [updated 2026, April 16. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
3. Román M, Sala M, De La Vega M, Baré M, Salas D, Ascunce N. Mammographic breast density: How it affects performance indicators in screening programmes? European Journal of Radiology. 2019;110:81–8.
4. Zackrisson S, Lång K, Rosso A, Johnson K, Dustler M, Förnvik D, et al. One-view breast tomosynthesis versus two-view mammography in the Malmö Breast Tomosynthesis Screening Trial (MBTST): A prospective, population-based, diagnostic accuracy study. The Lancet Oncology. 2018;19(11):1493–503.
5. Svahn TM, Chakraborty DP, Ikeda D, Zackrisson S, Do Y, Mattsson S, et al. Breast tomosynthesis and digital mammography: a comparison of diagnostic accuracy. British Journal of Radiology. 2012;85(1019):e1074–e82.
6. Lång K, Andersson I, Rosso A, Bengtsson E, Timberg P, Zackrisson S. Performance of breast tomosynthesis in screening mammography: Results from the Malmö Breast Tomosynthesis Screening Trial. European Radiology. 2016;26(11):3899–907.
7. Moger TA, Swanson JO, Holen Å S, Hanestad B, Hofvind S. Cost differences between digital tomosynthesis and standard digital mammography in a breast cancer screening programme: results from the To-Be trial in Norway. Eur J Health Econ. 2019;20(8):1261–9.
8. Abdullah P, Alabousi M, Ramadan S, Zawawi I, Zawawi M, Bhogadi Y, et al. Synthetic 2D Mammography Versus Standard 2D Digital Mammography: A Diagnostic Test Accuracy Systematic Review and Meta-Analysis. American Journal of Roentgenology. 2021;217(2):314–25.
9. Dustler M, Hellgren G, Timberg P. Comparing synthetic mammograms based on wide-angle digital breast tomosynthesis with digital mammograms. Journal of Medical Imaging. 2025;12(S1):S13011.
10. Dodelzon K, Simon K, Dou E, Levy AD, Michaels AY, Askin G, et al. Performance of 2D Synthetic Mammography Versus Digital Mammography in the Detection of Microcalcifications at Screening. American Journal of Roentgenology. 2020;214(6):1436–44.
11. Diagnostic Radiology Physics: A Handbook for Teachers and Students. First ed. Vienna: International Atomic Energy Agency (IAEA); 2014.
12. Gingold EL, Wu X, Barnes GT. Contrast and dose with Mo-Mo, Mo-Rh, and Rh-Rh target-filter combinations in mammography. Radiology. 1995;195(3):639–44.
13. Chikarmane S. Synthetic Mammography: Review of Benefits and Drawbacks in Clinical Use. Journal of Breast Imaging. 2022;4(2):124–34.
14. Tingberg A, Dahlblom V, Dustler M, Förnvik D, Johnson K, Timberg P, et al. Our journey toward implementation of digital breast tomosynthesis in breast cancer screening: the Malmö Breast Tomosynthesis Screening Project. Journal of Medical Imaging. 2024;12(S1):S13006.
15. IAEA. Diagnostic Radiology Physics: A Handbook for Teachers and Students. First ed. Vienna: International Atomic Energy Agency (IAEA); 2014.
16. Knoll GF. Radiation Detection and Measurement. 4 ed. Hoboken, NJ: John Wiley & Sons; 2010.

17. Jögi A, Johnson K, Wittgren S, Sundgren V, Tomic H, Olinder J, et al. Assessing Digital Breast Tomosynthesis Impact on Early Cancer Detection: Insights from Consecutive Screening. Radiology. 2024;312(1):e233417.
18. Olinder J, Förnvik D, Dahlblom V, Lu V, Åkesson A, Johnson K, et al. Assessing mammographic density change within individuals across screening rounds using deep learning–based software. Journal of Medical Imaging. 2025;12(S2):S22017.
19. Sechopoulos I. A review of breast tomosynthesis. Part II. Image reconstruction, processing and analysis, and advanced applications. Medical Physics. 2013;40(1):014302.
20. Sechopoulos I, Suryanarayanan S, Vedantham S, D'Orsi C, Karellas A. Computation of the glandular radiation dose in digital tomosynthesis of the breast. Med Phys. 2007;34(1):221–32.
21. Wu T, Stewart A, Stanton M, McCauley T, Phillips W, Kopans DB, et al. Tomographic mammography using a limited number of low-dose cone-beam projection images. Medical Physics. 2003;30(3):365–80.
22. Dahlblom V, Dustler M, Zackrisson S, Tingberg A. Workload reduction of digital breast tomosynthesis screening using artificial intelligence and synthetic mammography: a simulation study. Journal of Medical Imaging. 2025;12(S2):S22005.
23. Socialstyrelsen. Bröstcancer: Cancer i siffror 2023: Populärvetenskapliga fakta om cancer; 2023 [Available from: https://www.socialstyrelsen.se/statistik-och-data/statistik/alla-statistikamnen/cancer/.
24. Socialstyrelsen. Screening för bröstcancer: Socialstyrelsens rekommendation (Slutversion 2023). Stockholm; 2023. Report No.: 2023-5-8564.
25. Cancer IAfRo. Breast Cancer Screening: IARC Handbook of Cancer Prevention, Volume 15. Lyon: IARC; 2016.
26. Socialstyrelsen. Nationellt screeningprogram för bröstcancer. Stockholm; 2022.
27. samverkan Rci. Nationellt vårdprogram för bröstcancerscreening. Stockholm; 2022.
28. Healthineers S. MAMMOMAT B.brilliant [Available from: https://www.siemens-healthineers.com/se/mammography/digital-mammography/mammomat-bbrilliant.
29. Lillé S, Marshall W. Mammographic Imaging: A Practical Guide. Fourth Edition ed. Philadelphia: Wolters Kluwer Health; 2018. 600 p.
30. Sechopoulos I. A review of breast tomosynthesis. Part I. The image acquisition process. Med Phys. 2013;40(1):014301.
31. samverkan Rci. Nationellt vårdprogram för bröstcancerscreening. Stockholm; 2022.
32. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230–43.
33. Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. Seminars in Cancer Biology. 2021;72:214–25.
34. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019;111(9):916–22.
35. Dembrower K, Wåhlin E, Liu Y, Salim M, Smith K, Lindholm P, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. The Lancet Digital Health. 2020;2(9):e468–e74.

36. Gommers J, Hernström V, Josefsson V, Sartor H, Schmidt D, Hjelmgren A, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. The Lancet. 2026;407(10527):505–14.
37. Medical S. Transpara: Clinical performance and validation summary. Nijmegen, Netherlands; 2024.
38. Rodríguez-Ruiz A, Krupinski E, Mordang J-J, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology. 2019;290(2):305–14.
39. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861–74.
40. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. Bmj. 1994;308(6943):1552.
41. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
42. Lee SE, Han K, Kim EK. Application of artificial intelligence-based computer-assisted diagnosis on synthetic mammograms from breast tomosynthesis: comparison with digital mammograms. Eur Radiol. 2021;31(9):6929–37.