Check for updates

Received: 12 July 2022 | Revised: 6 April 2023 | Accepted: 11 April 2023

DOI: 10.1002/mp.16439

RESEARCH ARTICLE

MEDICAL PHYSICS

Deep learning denoising of digital breast tomosynthesis: Observer performance study of the effect on detection of microcalcifications in breast phantom images

Heang-Ping Chan | **Mark A. Helvie** | **Mingjie Gao** | **Lubomir Hadjiiski** | **Chuan Zhou** | **Kim Garver** | **Katherine A. Klein** | **Carol McLaughlin** | **Rebecca Oudsema** | **W. Tania Rahman** | **Marilyn A. Roubidoux**

Department of Radiology, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Heang-Ping Chan, Department of Radiology, University of Michigan, 1500 E. Medical Center Drive, Med Inn Building C477, Ann Arbor, MI 48109-5842, USA.
Email: chanhp@umich.edu

Funding information
National Institutes of Health, Grant/Award Number: RO1 CA214981

Abstract

Background: The noise in digital breast tomosynthesis (DBT) includes x-ray quantum noise and detector readout noise. The total radiation dose of a DBT scan is kept at about the level of a digital mammogram but the detector noise is increased due to acquisition of multiple projections. The high noise can degrade the detectability of subtle lesions, specifically microcalcifications (MCs).

Purpose: We previously developed a deep-learning-based denoiser to improve the image quality of DBT. In the current study, we conducted an observer performance study with breast radiologists to investigate the feasibility of using deep-learning-based denoising to improve the detection of MCs in DBT.

Methods: We have a modular breast phantom set containing seven 1-cm-thick heterogeneous 50% adipose/50% fibroglandular slabs custom-made by CIRS, Inc. (Norfolk, VA). We made six 5-cm-thick breast phantoms embedded with 144 simulated MC clusters of four nominal speck sizes (0.125–0.150, 0.150–0.180, 0.180–0.212, 0.212–0.250 mm) at random locations. The phantoms were imaged with a GE Pristina DBT system using the automatic standard (STD) mode. The phantoms were also imaged with the STD+ mode that increased the average glandular dose by 54% to be used as a reference condition for comparison of radiologists' reading. Our previously trained and validated denoiser was deployed to the STD images to obtain a denoised DBT set (dnSTD). Seven breast radiologists participated as readers to detect the MCs in the DBT volumes of the six phantoms under the three conditions (STD, STD+, dnSTD), totaling 18 DBT volumes. Each radiologist read all the 18 DBT volumes sequentially, which were arranged in a different order for each reader in a counter-balanced manner to minimize any potential reading order effects. They marked the location of each detected MC cluster and provided a conspicuity rating and their confidence level for the perceived cluster. The visual grading characteristics (VGC) analysis was used to compare the conspicuity ratings and the confidence levels of the radiologists for the detection of MCs.

Results: The average sensitivities over all MC speck sizes were 65.3%, 73.2%, and 72.3%, respectively, for the radiologists reading the STD, dnSTD, and STD+ volumes. The sensitivity for dnSTD was significantly higher than that for STD ($p < 0.005$, two-tailed Wilcoxon signed rank test) and comparable to that for STD+. The average false positive rates were 3.9 ± 4.6, 2.8 ± 3.7, and 2.7 ± 3.9

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2023 The Authors. *Medical Physics* published by Wiley Periodicals LLC on behalf of American Association of Physicists in Medicine.

marks per DBT volume, respectively, for reading the STD, dnSTD, and STD+ images but the difference between dnSTD and STD or STD+ did not reach statistical significance. The overall conspicuity ratings and confidence levels by VGC analysis for dnSTD were significantly higher than those for both STD and STD+ ($p \leq 0.001$). The critical alpha value for significance was adjusted to be 0.025 with Bonferroni correction.

Conclusions: This observer study using breast phantom images showed that deep-learning-based denoising has the potential to improve the detection of MCs in noisy DBT images and increase radiologists' confidence in differentiating noise from MCs without increasing radiation dose. Further studies are needed to evaluate the generalizability of these results to the wide range of DBTs from human subjects and patient populations in clinical settings.

KEYWORDS
deep learning denoising, digital breast tomosynthesis, microcalcifications, observer study

1 | INTRODUCTION

Digital breast tomosynthesis (DBT) has become a commonly available breast cancer screening modality since it was approved for clinical use about a decade ago. A meta-analysis including 17 studies showed that the cancer detection rates increased by using DBT in combination with digital mammogram (DM), while the recall rates reduced in the United States but increased in Europe, as compared to DM alone in screening.[1] Another meta-analysis including 11 studies showed that the improved cancer detection rates could mainly be attributed to invasive cancers. The detection of in situ cancers on average did not increase significantly and about half of the reviewed studies actually observed a decrease.[2] Bahl et al.[3] showed that in situ cancers contributed to a substantially smaller proportion of the screen-detected cancers when screening with DBT + DM compared to DM alone and the reduction in the proportion of in situ cancers sustained over the 5 years that they analyzed.[4]

Ductal carcinoma in situ (DCIS) often manifests as microcalcifications (MCs) alone and the detection of some subtle invasive cancers is aided when there are associated MCs. Detection of MCs in DBT is more challenging than in DM because many more images have to be read in a DBT volume than in a two-dimensional (2D) DM image for each breast compression view. The reported detection sensitivity of MCs in DBT varied, probably due to the different physical characteristics among the manufacturers' DBT systems, factors that affect the image quality, and importantly, the different degrees of subtlety of the MC cases that might be evaluated in the studies.[5–12] Using DM in combination with DBT would allow radiologists to maintain their sensitivity for detecting MCs, but it increases the radiation to the patient. Replacing DM with a synthetic DM-like mammogram (SM) can eliminate the additional dose. The SM technology generally is implemented with computer-aided detection and image processing techniques to enhance suspicious lesions.[13–16] The noise in DBT can reduce the detectability of MCs and also contribute to pseudocalcifications on the SM due to over-enhancement.[17] The detection of MCs on DM and SM was compared previously.[17–21] Other studies showed that DCIS detection rate in DBT + SM was lower compared to DBT + DM or DM alone.[18,22–24] A two-center study showed that the screen-detected DCIS rate was lower with DBT + SM relative to DBT + DM in one center but higher in the other, and the overall rate did not show significant difference.[25] A multi-center, multi-reader study[18] and a recent study of community-based screening[26] showed that the percentages of DCIS and cancers manifested as MCs among the detected cancers were significantly lower with DBT + SM than with DM alone. Although the reduction in DCIS detection often did not reach statistical significance in many of these studies, the overall trend may not be negligible.

Conventional approach to alleviating the noise problem is to use imaging techniques at higher dose[27] or reduce the noise with smoothing filtering. Several studies investigated the use of conventional filters to reduce noise in DBT images and reported various degrees of success.[28–33] The recent advances of machine learning technologies spur the development of deep learning (DL)-based techniques for improving the image quality of medical images. Some studies showed that DL-based denoising could effectively reduce noise in low dose computed tomography (CT),[34–36] while other studies did not find DL-based denoising to be effective in some nuclear medicine applications.[37–39] The image characteristics of MCs in DBT are very different from lesions in CT or nuclear medicine images. The major challenge is that subtle MCs cannot be easily distinguished from noise. Several studies have attempted to develop DL-based denoising for DBT.[40–45] Badal et al. evaluated three DCNN models for denoising DBT projections and reported that the detectability of mass was not affected but the slight blurring of the images by the DCNNs reduced the detectability of MC cluster.[46] We have

24734209, 2023, 10, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16439 by University Of Michigan Library, Wiley Online Library on [15/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

FIGURE 1 Modular breast phantom set with 1-cm-thick slabs of heterogeneous distribution of breast-tissue-equivalent material simulating a nominal 50% adipose/50% fibroglandular composition (custom-made by CIRS, Inc.).

investigated a DL-based approach to denoising reconstructed DBT images and demonstrated its feasibility of enhancing MCs based on the contrast-to-noise ratio and sharpness measures and a task-based detectability index.[47] However, these image quality measures only analyze individual MCs without taking into account the detectability of MC clusters in a search task. The goal of this study is to evaluate the effects of DL-based denoising on radiologists' detection of MC clusters in breast phantom DBTs and, as a secondary analysis, examine the dependence of the effects of the denoiser on MC speck sizes.

2 | MATERIALS AND METHODS

2.1 | Breast phantoms

We have a modular breast phantom set containing seven 1-cm-thick breast-tissue-equivalent slabs of nominal 50% adipose/50% fibroglandular composition custom-made by CIRS, Inc. (Norfolk, VA). The semicircular shaped slabs were approximately 20 cm in diameter and had random heterogeneous patterns to simulate breast tissue structures, as shown in Figure 1. We constructed six different 5-cm-thick breast phantoms by stacking five slabs from the seven slabs in different orders such that the order of any two adjacent slabs in one phantom would not be repeated in another phantom. Glass beads (Whitehouse Scientific Ltd.) of four nominal speck size ranges (0.125–0.150, 0.150–0.180, 0.180–0.212, and 0.212–0.250 mm) were used to form simulated MC clusters for each speck size. These glass beads were chosen because glass beads are used in the ACR phantom for digital mammography (model 086, CIRS, Inc.) to simulate calcification specks and the four chosen nominal size ranges for our phantoms covered the four smaller sizes of the six speck groups in the ACR phantom. Twenty-four clusters, six of each speck size were mixed and sandwiched between the slabs at random locations in each phantom. Table 1 shows the number of breast phantoms and the number of simulated clusters of each speck size in each phantom.

2.2 | DBT imaging

The six breast phantoms were imaged with a GE Pristina DBT system (GE Healthcare, Waukesha, WI). Each phantom was compressed to an average thickness of 50.3 mm (range: 50.0–50.9 mm) at a compression force of about 8 daN. The automatic exposure modes, standard (STD) and STD+, available in the DBT system were used to consecutively image the phantom under the same compression. The STD mode is the routine clinical technique for patient imaging. The STD+ mode was included in this study to provide a reference condition for comparison of radiologists' reading whether the denoised STD images could match those acquired at a clinically practical higher dose level. The system used a fixed target/filter (Rh/Ag) and 34 kV technique for breasts thicker than about 40 mm. For the STD mode, the average mAs was 33.1 (range: 31.5–34.2 mAs) and the average glandular dose (AGD) was 1.34 mGy (range: 1.31–1.36 mGy). For the STD+ mode, the average mAs was 54.2 (range: 48.9–57.2 mAs) and the AGD was 2.07 mGy (range: 1.91–2.15 mGy). The AGD of STD+ was about 54% higher than that of the STD. All DBTs were reconstructed by the GE proprietary reconstruction software provided with the Pristina system at a pixel size of 0.1 mm × 0.1 mm and 1-mm spacing.

2.3 | Deep learning-based denoiser

We previously developed a deep convolution neural network (DCNN)-based 2D denoiser for reducing noise in reconstructed DBT images. Details of our study including the methodology and the experiments that we conducted for training, validation, and testing of the denoise have been published.[47] A brief summary is given below.

As illustrated in Figure 2, the denoiser is designed based on a generative adversarial network (GAN) that is composed of a generator and a discriminator, both are DCNNs with trainable weights. We therefore refer to our denoising GAN as DNGAN. The DCNN architecture of the generator contained 10 convolutional layers, each with 32 filters of 3 × 3 kernels, and used rectified linear units (ReLU) between the layers. The discriminator used a VGG-net backbone but a reduced number of downsampling blocks to adapt to the small input patch size used for training the denoiser. It contained a total of six convolutional layers, three 2 × 2 max pooling

24734209, 2023, 10, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16439 by University Of Michigan Library, Wiley Online Library on [15/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 The test DBT volumes (6 volumes × 3 reading conditions) used in the observer study

Reading conditions		STD mode	dnSTD (denoised STD)	STD+ mode (reference)
Number of DBT volumes		6	6	6
Number of simulated MC clusters of each speck size embedded in each of the DBT volumes	0.125–0.150 mm	6	6	6
	0.150–0.180 mm	6	6	6
	0.180–0.212 mm	6	6	6
	0.212–0.250 mm	6	6	6
Total number of simulated MC clusters embedded in the set of six DBT volumes		144	144	144

Note: Twenty-four simulated MC clusters (=4 × 6 of each speck size) were mixed and sandwiched between the slabs at random locations in each phantom, resulting in a total of 144 (=24 × 6) simulated MC clusters under each reading condition to be detected by radiologists. Two other DBT volumes used in the observer training session is not listed here.

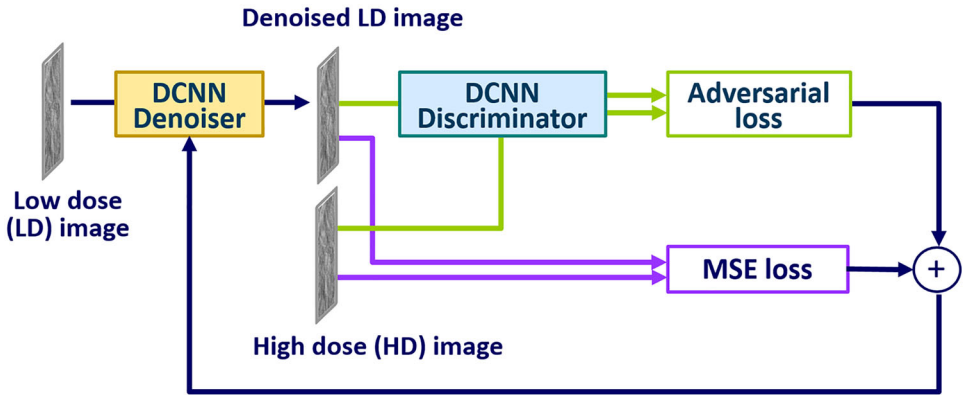

FIGURE 2 A schematic of the Wasserstein GAN-based adversarial training of the denoiser DNGAN. Note that the discriminator and the target high dose (HD) images are needed only during training. After training, the DNGAN with the frozen trained weights can standalone and be deployed to reduce noise of an input low dose (LD) image.

layers, and one fully connected layer. The number of filters in each convolutional layer in the discriminator ranged from 32 to 128, each of 3 × 3 kernel.

The DNGAN training required a large set of low-dose/high-dose (LD/HD) image pairs as input and target output images.[47] We used image pairs having a patch size of 32 × 32 pixels. The denoiser training was guided by minimizing a training loss function, which was composed of the mean squared error (MSE) loss, L_{MSE}, and the adversarial loss, L_{adv}.

$$\underset{G}{\operatorname{argmin}}\, L_{MSE}(G) + \lambda_{adv} \cdot L_{adv}(G)$$

where G was the denoiser. The MSE loss compared the pixel-wise difference between the denoised image patches and the corresponding HD target image patches. The adversarial loss was derived from training the discriminator that assessed the similarity between the distributions of the denoised low-dose images and the target high-dose images. The MSE loss contributed to image smoothness while the adversarial loss contributed to preserving the high frequency image texture. The Wasserstein GAN-based adversarial training in which the generator and the discriminator were trained alternately was adopted to constrain the degree of smoothing and maintain the sharpness of the denoised DBT images.[47] The parameter, λ_{adv}, that weighted the two training loss terms was chosen experimentally to further balance between noise smoothing and image structure fidelity. After training and validation, the trained generator with its frozen weights can be deployed as a denoiser to reduce noise in an input noisy DBT image while maintaining its structural details. As the denoiser is fully convolutional, it can be applied to a full size DBT image during deployment. The details of the specific DCNN structures, the adversarial training and parameter selection process can be found in our previous paper.[47]

The DNGAN training requires pairs of LD/HD images but pairs of LD/HD human DBT images under the same compression are not available. As described previously,[47] we conducted an extensive study to train the DNGAN denoiser using DBT images of digital breast phantoms or physical phantoms and validate the robustness of the trained denoisers by deploying them to

24734209, 2023, 10, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16439 by University Of Michigan Library, Wiley Online Library on [15/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

independent digital phantom, physical phantom, and human subject DBT images, as summarized below.

For training with digital phantom images, we generated twenty-five heterogeneous dense (34% glandular volume fraction) 4.5-cm-thick digital phantoms using the VICTRE breast model.[48] We then used the CATSim (GE Global Research) simulation packages to model the Pristina DBT system in terms of target/filter and scan geometry. Pairs of LD/HD DBT images of each digital breast phantom were acquired with the virtual Pristina system over a wide range of dose levels (LD at 24 mAs, HD at 72, 120, 360 mAs and noiseless). All simulated DBTs were reconstructed with the simultaneous algebraic reconstruction technique (SART).[49] About 200 000 pairs of 32 × 32-pixel 2D image patches were extracted randomly from the set of reconstructed DBT volumes for each LD/HD condition to study the effects of the dose level of the HD target output on denoiser performance.

For training with physical phantom images, we acquired LD/HD image pairs of eight physical breast phantoms with a clinical GE Pristina system. The LD images were acquired at the STD mode (Rh/Ag 34 kVp, 31.4 mAs on average) and the HD images at about four times STD (manually set at Rh/Ag 34 kVp, 125 mAs). All DBT volumes were reconstructed with the GE reconstruction algorithm on the Pristina system and also with the SART. A total of 400 000 LD/HD pairs of 32 × 32-pixel 2D image patches were extracted randomly from each set of reconstructed DBT volumes for training.

For validation and testing, the denoiser trained with the training set from each of the different paired LD/HD image sources and reconstruction algorithms, described above, was deployed to an independent physical phantom DBT volume containing a total of 236 simulated MCs of size 0.150–0.180 mm, 227 of 0.180–0.212 mm, and 159 of 0.212–0.250 mm. The coordinates of the individual MCs were manually marked for analysis. The contrast-to-noise ratio (CNR), the full width at half maximum, and a task-based detectability index (d′) from the nonprewhitening matched filter model observer with eye filter of the individual MCs were quantified and then averaged over all MCs of each size range as performance measures. The corresponding performance measures of the various trained denoisers were compared among the LD, HD, and denoised LD images to validate the generalizability of the DNGAN denoiser to unseen DBT images. In addition, the CNRs of 301 real MCs in human subject DBTs with and without DL denoising were compared. The noise power spectra of the LD, HD, and denoised LD images were compared on digital phantom DBT images to evaluate the effects of the denoisers trained with different weights of the adversarial loss, L_{adv}, and at various dose levels of the target HD images on image texture and high frequency structures. The image quality of the denoised DBT images for physical phantoms and human subjects were also compared visually for the different denoisers. The results of the previous study indicated that either of the training approaches using digital phantom or physical phantom images could train a DNGAN that can effectively reduce noise while preserving the structural details in noisy DBT images of both phantoms and human subjects.[47]

In the current study, we used the DNGAN trained with physical phantom images acquired with a GE Pristina DBT system and validated in the previous study,[47] as described above. The validated DNGAN denoiser was applied, without retraining, to a new test set of reconstructed DBT images acquired specifically for the observer study, as described in Section 2.2. Note that the denoiser, once trained, does not require the HD images when it is applied to unseen test images. The STD+ image set was included only as a reference condition for comparison of radiologists' reading performance in the current study.

2.4 | Observer study

We conducted a multi-reader multi-case observer performance study to compare the detection of MC clusters in DBT images obtained from three conditions: STD images, STD+ images, and DNGAN-denoised images (dnSTD). Each of the conditions had six DBT volumes, totaling 18 volumes (=3 conditions × 6 volumes per condition). Seven MQSA-certified radiologists experienced in mammography (2–34 years, median: 13 years) participated as observers. Each observer read all 18 DBT volumes sequentially that were arranged in an order that was different from any other observers' sequences. The three conditions of the same breast phantom were mixed in the sequence but separated as far apart as possible. No washout period was needed because there were too many background patterns and MC locations to memorize. The observer was blinded to the condition of the DBT volume being read. The DBT volumes was ordered in a counter-balanced manner such that the DBT of a given condition being read first, second, or last was balanced when averaged over all volumes and all readers to avoid biases due to reading order effects.[50]

The DBT volume was displayed on a 21″ 5M-pixel (2048 × 2560) LCD display monitor (model EIZO SMD 21500 D Contrast ratio 800:1, maximum luminance 750 cd/mm^2) calibrated with the DICOM grayscale standard display function. An in-house developed graphical user interface was used to display the DBT volume, which allowed the observer to adjust the contrast and brightness, scroll through the volume, and pan and zoom the images as needed. The observer could mark a detected MC cluster with a box, provide a conspicuity rating of the cluster (10-point scale: 1 = subtle, 10 = obvious) and their confidence level (10-point scale: 1 = low, 10 = high) that the marked region contained an MC cluster. Multiple regions could be marked in a single DBT volume and each with its own ratings. No time limit was

24734209, 2023, 10, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16439 by University Of Michigan Library, Wiley Online Library on [15/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

imposed for the reading. Each observer could read at their own pace and search for MCs in their preferred way. They were asked to use the same search method and keep their own relative rating scales consistently over the entire reading experiment. The user interface automatically recorded the time that a detected location was marked so that the average time per mark could be estimated for each condition and each observer. The observers were free to separate the reading into multiple sessions and all observers finished the entire set in two or three sessions.

Each observer underwent a training session before reading the test DBT volumes. They were shown DBT volumes of training phantoms, different from the test phantoms, on the display monitor and marked the detected MC clusters to be familiarized with the functions of the graphical user interface and the rating scales. After they marked the potential MC clusters in a training volume, a corresponding high dose volume superimposed with the ground truth locations was shown side by side for the observers to review their true positive, false positive, and false negative detections and learned the appearance of the MC clusters of various speck sizes. They were informed that each DBT volume contained a large number of MC clusters of various degrees of subtlety but the number of clusters in a given volume or at a given depth was not disclosed. The true locations of the clusters in the DBT volumes of the test set were never shown to the observers.

2.5 | Data analysis

We estimated radiologists' sensitivity for detecting the simulated MC clusters of each speck size in all DBT volumes. The differences in the average detection sensitivity with all speck sizes together were compared between two pairs of the image conditions, dnSTD versus STD and dnSTD versus STD+, read by the radiologists. The false positive (FP) rate, defined as the average number of FP marks per DBT volume, was calculated for each radiologist and the average FP rate over all radiologists derived. The statistical significance of the difference between the paired conditions was estimated by the two-tailed Wilcoxon's signed rank test.

For the confidence levels and conspicuity ratings, we performed the multiple-reader multiple-case visual grading characteristics (VGC) analysis to compare the paired conditions.[51–53] The VGC analysis is a non-parametric rank-invariant statistical method for comparing image quality of two conditions (or modalities).[51,52] The differences in the rating data under the two conditions by the observers for a specific image quality indicator such as the conspicuity of a structure are analyzed in a way similar to differentiating the two classes in the receiver operating characteristic analysis, resulting in a VGC curve. The area under the VGC curve (AUC_{vgc}) is a measure of the difference between the two conditions. An AUC_{vgc} of 0.5 shows that there is no difference in the specific image quality indicator being rated under the two conditions. The difference is considered statistically significant if the 95% confidence interval for the estimated AUC_{vgc} does not cover 0.5. We performed the VGC analysis using the fully-crossed, multiple-reader multiple-case software developed and validated by Båth et al.[52,53]

With the comparison of two paired reading conditions (dnSTD vs. STD and dnSTD vs. STD+), the critical alpha level for estimation of significance would be adjusted by a factor of 2 to 0.025 according to the Bonferroni method. A p-value of less than 0.025 was considered statistically significant.

As a secondary analysis, we compared the average detection sensitivity, the MC conspicuity rating and the radiologists' confidence level for each of the speck sizes. The purpose of this analysis was to examine the trends, if any, of the effects of speck size on the DL-based denoiser performance, which may provide useful information to guide future research efforts. No statistical significance test was applied because the subgroup analysis was not intended to influence the main goal of the current study.

3 | RESULTS

Figure 3 shows a DBT image of one of the breast phantoms. Figure 4 shows close-up views of a simulated MC cluster of each nominal speck size for the three image conditions. The dnSTD images show that the parenchymal structures are well preserved and less noisy than those in the STD and STD+ images.

The radiologists' sensitivities of detecting the simulated grouped MCs averaged over all speck sizes were 65.3%, 73.2%, and 72.3%, respectively, for the STD, dnSTD, and STD+ conditions. The average sensitivity for dnSTD was significantly higher than that for STD ($p < 0.005$, two-tailed Wilcoxon signed rank test), and was comparable to that of STD+. The average FP rates were 3.9 ± 4.6, 2.8 ± 3.7, and 2.7 ± 3.9 marks per DBT volume, respectively, for reading the STD, dnSTD, and STD+ images, but none of the differences reach statistical significance.

Table 2 shows the AUC_{vgc} values from the VGC analysis. The comparison indicated that the radiologists' conspicuity ratings and confidence levels for MC detection in the dnSTD images were significantly higher than those in the STD+ and STD images ($p \leq 0.001$).

The average sensitivities of detecting the simulated MC clusters of each speck size for the seven radiologists are shown in Table 3. Figure 5(a) compares the detection sensitivities averaged over the seven radiologists for each speck size under the three image conditions. The average sensitivities for the dnSTD and

24734209, 2023, 10, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16439 by University Of Michigan Library, Wiley Online Library on [15/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

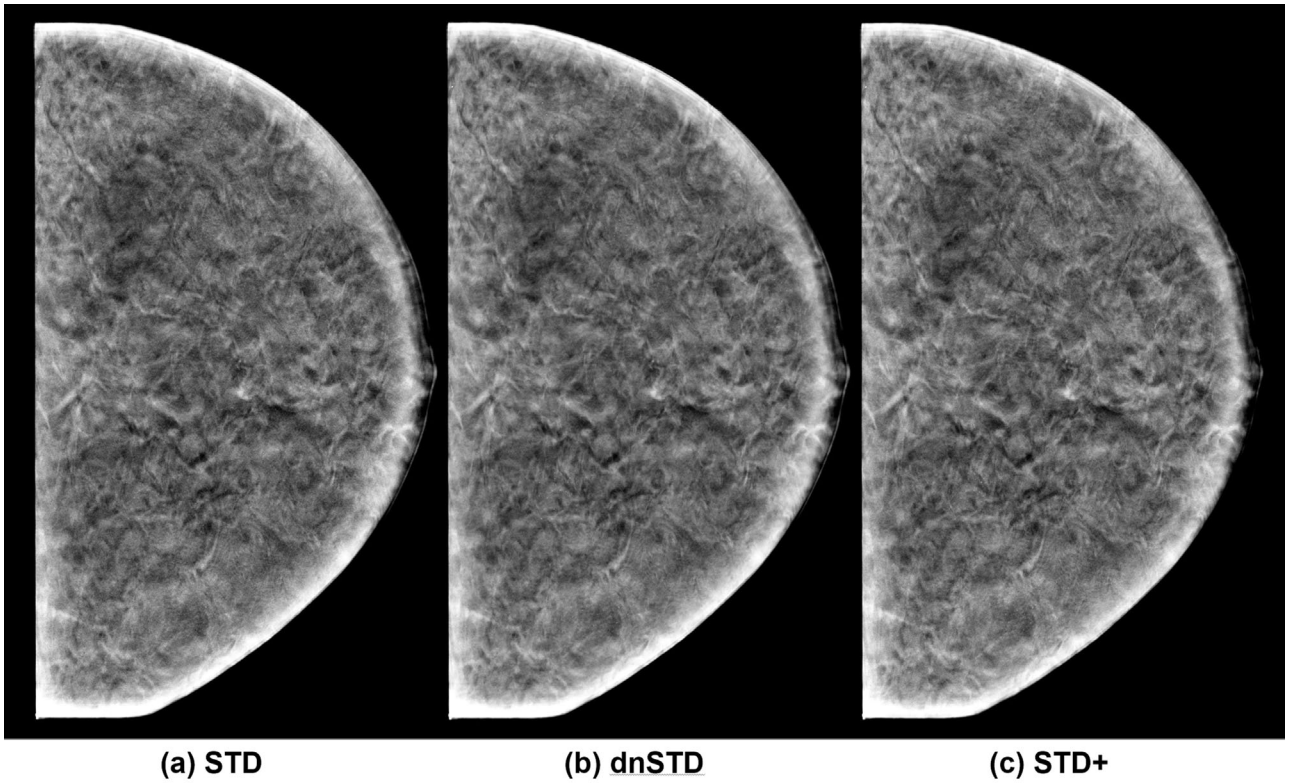

FIGURE 3 Example of a digital breast tomosynthesis (DBT) image of a breast phantom with simulated grouped microcalcifications for the three conditions: (a) STD, (b) denoised STD, (c) STD+.

TABLE 2 Multiple-reader multiple-case visual grading characteristics (VGC) analysis comparing radiologists' conspicuity ratings and confidence levels for the detection of the simulated grouped MCs of all speck sizes between pairs of the image conditions.

Comparison	dnSTD > STD	dnSTD > STD+
Conspicuity rating	0.646 (0.612, 0.684)	0.542 (0.513, 0.571)
p-value	<0.0001*	<0.0001*
Confidence level	0.643 (0.610, 0.679)	0.542 (0.512, 0.571)
p-value	<0.0001*	0.001*

Note: The conspicuity ratings and confidence ratings for dnSTD are significantly higher than those for STD+ and STD. The values shown are AUC_{vgc} (95% confidence intervals). The *p*-value shows the statistical significance of the difference from $AUC_{vgc} = 0.5$. The asterisk * indicates statistical significance at an adjusted critical alpha value of 0.025.

STD+ images were consistently higher than those for the STD images.

The average conspicuity ratings and confidence levels by the radiologists for each speck size are shown in Tables 4 and 5, respectively. Figure 5(b) compares the average conspicuity ratings under the three image conditions. The trends of the average conspicuity ratings and the confidence levels are consistent with that observed for the sensitivity. The clusters in both the dnSTD and the STD+ images had higher conspicuity and confidence ratings than those in the STD images. In addition, the clusters of three of the four speck sizes in the dnSTD images had higher average conspicuity and confidence ratings than those in the STD+ images, which was also observed for the average sensitivities. The relative quality of the dnSTD, STD+, and STD images can be seen in the examples in Figure 4.

The average reading time per mark reading the dnSTD images was reduced slightly by 7.5% compared to reading the STD images, and was essentially the same as that reading the STD+ images.

4 | DISCUSSION

This study indicated that the deep-learning-based denoiser could reduce the noise in DBT images acquired with a STD technique such that the detection sensitivity of MCs in dnSTD could increase significantly from that in STD to a level comparable to that in STD+. The increased sensitivity in the dnSTD images could be attributed to the improved contrast-to-noise ratio of MCs and therefore their conspicuity in the DBT images.[47] More importantly, the higher conspicuity also significantly increased the radiologists' confidence level in distinguishing grouped MCs from noise. The average FP detection rate in dnSTD was reduced substantially, although not statistically significant, by about 28% (=2.8/3.9) relative to STD and became comparable to that in STD+ (2.8 vs. 2.7). The improvement in the

24734209, 2023, 10, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16439 by University Of Michigan Library, Wiley Online Library on [15/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

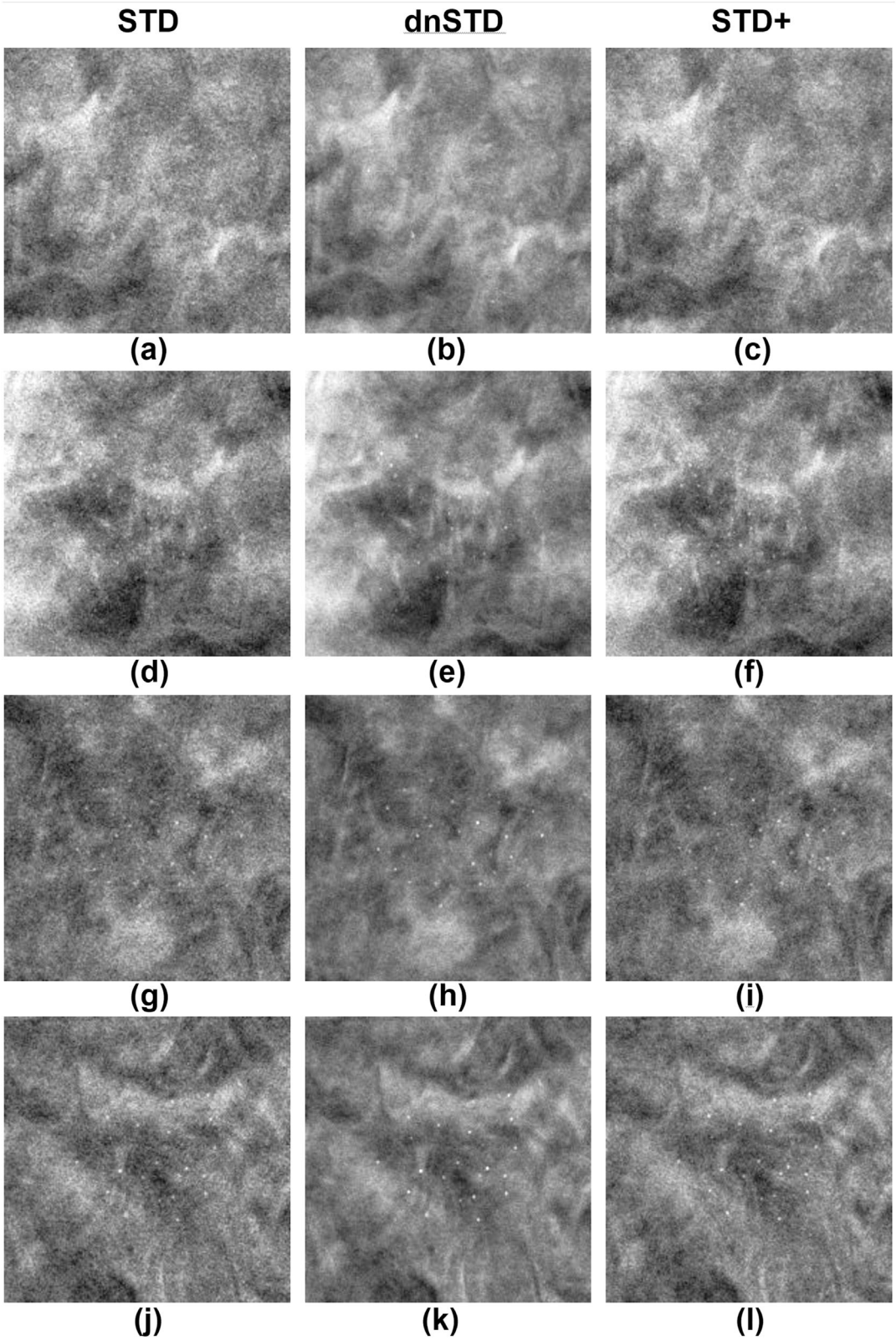

FIGURE 4 Examples of simulated grouped microcalcifications of four speck sizes: (a)–(c) 0.125–0.150 mm, (d)–(f) 0.150—0.180 mm, (g)–(i) 0.180–0.212 mm, (j)–(l) 0.212–0.250 mm. The image conditions are: (left column) STD, (middle column) denoised STD, (right column) STD+. Each of the image in (a)–(l) is 20 mm × 20 mm in size.

24734209, 2023, 10, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16439 by University Of Michigan Library, Wiley Online Library on [15/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 Radiologists' sensitivity of detecting the simulated grouped MCs for each nominal speck size

Speck size	0.125–0.150 mm			0.150–0.180 mm			0.180–0.212 mm			0.212–0.250 mm			Average false positives per volume		
	STD	dnSTD	STD+	STD	dnSTD	STD+	STD	dnSTD	STD+	STD	dnSTD	STD+	STD	dnSTD	STD+
R1	5.6%	2.8%	8.3%	50.0%	75.0%	83.3%	94.4%	97.2%	97.2%	91.7%	100%	100%	0.83	0.50	0.17
R2	0.0%	13.9%	0.0%	75.0%	88.9%	91.7%	91.7%	100%	100%	94.4%	100%	100%	0.83	1.17	0.33
R3	8.3%	11.1%	11.1%	47.2%	77.8%	77.8%	69.4%	86.1%	86.1%	77.8%	94.4%	86.1%	7.00	2.83	5.33
R4	5.6%	25.0%	13.9%	86.1%	94.4%	100%	100%	94.4%	100%	100%	97.2%	100%	0.33	1.33	0.17
R5	22.2%	44.4%	25.0%	91.7%	94.4%	100%	97.2%	97.2%	94.4%	94.4%	100%	100%	12.83	11.00	10.50
R6	16.7%	11.1%	13.9%	94.4%	94.4%	100%	100%	100%	100%	100%	100%	100%	1.00	1.33	0.50
R7	2.8%	5.6%	0.0%	44.4%	55.6%	55.6%	80.6%	91.7%	83.3%	86.1%	97.2%	97.2%	4.17	1.33	1.83
Mean	8.7%	16.3%	10.3%	69.8%	82.9%	86.9%	90.5%	95.2%	94.4%	92.1%	98.4%	97.6%	3.86	2.79	2.69
Std Dev	7.9%	14.3%	8.7%	22.1%	14.6%	16.4%	11.4%	5.0%	7.0%	7.9%	2.2%	5.2%	4.64	3.69	3.91

Note: The mean of dnSTD was higher than that of STD or STD+ for all paired comparisons except for dnSTD < STD+ at speck size of 0.150–0.180 mm. Statistical significance was not estimated for this secondary subgroup analysis.

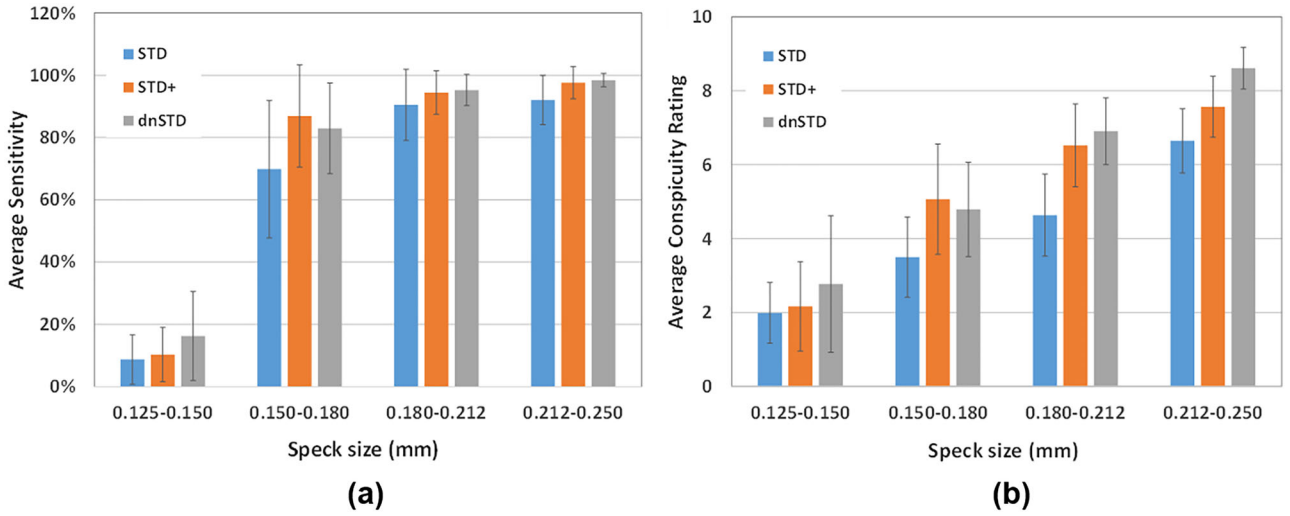

FIGURE 5 (a) Average sensitivities for detecting the simulated microcalcification (MC) clusters and (b) average conspicuity ratings for the detected clusters of each speck size. The error bars represent one standard deviation. In each group of bars, left to right: STD, STD+, dnSTD.

TABLE 4 Radiologists' conspicuity ratings (10-point scale: 1 = low, 10 = high) on the detected simulated grouped MCs for each nominal speck size

Speck size	0.125–0.150 mm			0.150–0.180 mm			0.180–0.212 mm			0.212–0.250 mm		
	STD	dnSTD	STD+	STD	dnSTD	STD+	STD	dnSTD	STD+	STD	dnSTD	STD+
R1	2.50	6.00	3.33	4.11	5.74	5.83	5.18	7.03	7.03	6.58	8.39	7.78
R2	1.00	1.00	1.00	2.33	3.69	3.73	3.58	6.81	5.97	7.12	8.75	7.64
R3	2.67	2.50	2.75	4.24	3.86	5.54	4.68	6.23	6.68	7.14	8.50	7.52
R4	1.00	1.11	1.00	2.06	4.09	3.33	3.28	6.71	5.50	5.50	8.43	7.08
R5	3.13	2.56	3.89	5.06	7.03	7.47	6.20	8.66	8.38	7.41	9.61	8.89
R6	1.67	1.75	2.20	3.00	3.82	3.83	3.86	5.78	5.06	5.39	7.75	6.17
R7	2.00	4.50	1.00	3.69	5.30	5.75	5.69	7.15	7.07	7.39	8.86	7.91
Mean	1.99	2.77	2.17	3.50	4.79	5.07	4.64	6.91	6.53	6.65	8.61	7.57
Std Dev	0.82	1.85	1.21	1.08	1.27	1.49	1.11	0.91	1.12	0.87	0.57	0.83

Note: The mean of dnSTD was higher than that of STD or STD+ for all paired comparisons except for dnSTD < STD+ at speck size of 0.150–0.180 mm. Statistical significance was not estimated for this secondary subgroup analysis.

24734209, 2023, 10, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16439 by University Of Michigan Library, Wiley Online Library on [15/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5 Radiologists' confidence levels (10-point scale: 1 = low, 10 = high) on the detected simulated grouped MCs for each nominal speck size

Speck size	0.125–0.150 mm			0.150–0.180 mm			0.180–0.212 mm			0.212–0.250 mm		
	STD	dnSTD	STD+	STD	dnSTD	STD+	STD	dnSTD	STD+	STD	dnSTD	STD+
R1	3.00	6.00	4.33	4.61	6.41	6.43	5.97	7.03	7.03	6.97	8.06	7.64
R2	1.00	1.40	1.00	2.63	4.25	4.55	4.36	6.81	6.44	7.56	8.75	7.83
R3	2.67	3.00	2.75	4.59	4.21	5.50	5.20	6.58	6.68	7.11	8.18	7.52
R4	1.00	1.11	1.00	1.77	3.47	2.75	2.81	5.79	4.58	4.67	7.74	6.08
R5	3.50	3.06	3.78	5.21	6.97	7.31	6.43	8.54	8.06	7.44	9.58	8.89
R6	1.67	1.75	2.00	2.97	3.76	3.94	3.94	5.64	5.08	5.50	7.53	5.94
R7	2.00	4.00	1.00	3.44	5.20	5.30	5.24	7.09	6.80	7.16	8.83	7.80
Mean	2.12	2.90	2.27	3.60	4.90	5.11	4.85	6.78	6.38	6.63	8.38	7.39
Std Dev	0.98	1.71	1.39	1.24	1.35	1.53	1.24	0.96	1.18	1.10	0.71	1.04

Note: The mean of dnSTD was higher than that of STD or STD+ for all paired comparisons except for dnSTD < STD+ at speck size of 0.150–0.180 mm. Statistical significance was not estimated for this secondary subgroup analysis.

detection performance was achieved without an increase in the radiation dose or reading time.

The DL denoiser used in the observer study was trained with physical phantom image pairs acquired in STD mode as LD and four times the dose of STD as HD target in our previous study.[47] We included STD+ mode in radiologists' reading as a reference condition to evaluate whether the performance in dnSTD could match that from a clinically practical STD+ dose level, which was only 54% higher than that of STD. The results showed that radiologists' sensitivity in reading the dnSTD images was only marginally better than that in reading STD+ images although the MC conspicuity ratings in dnSTD were significantly higher than those in STD+. This may reflect the fact that human's visual search task is not only affected by the conspicuity of the target signals but also by other factors involved in signal detection.

The detection of MCs in DBT images is a time-consuming but critically important task for radiologists because subtle MCs are difficult to distinguish from noise but may be the only finding of malignancy. Studies of DBT for improving breast cancer detection predominantly focused on invasive cancers most of which manifested as masses or architectural distortion. However, some invasive cancers are mammographically detected by MCs alone or subtle soft tissue findings associated with MCs. DCIS manifests primarily as MCs and is pre-invasive. An ACRIN study found that a substantial proportion of DCIS (7.5% and 13.4%, respectively) can progress to invasive cancer even after surgical excision regardless whether it was low grade or high grade at diagnosis.[54] Another study showed that 26% of the low-grade and 31% of the high-grade DCIS in a surveillance group (excision with margin clearance < 1 mm) recurred as invasive cancer within 10 years.[55] In a population-based study, the screen-detected DCIS rate was found to be significantly associated with a reduction in the screen-detected invasive interval cancer rate in subsequent years.[56] These studies indicate that early detection of DCIS is important to reduce the chance that DCIS progresses into invasive cancer. Although there is concern of overdiagnosis, primarily among older women, early detection is still the key step that will offer the patient and the physician opportunity to make informed decision on management options.

We used breast phantoms in this preliminary study. The advantages are that we could acquire the higher-dose STD+ image as a reference condition in the observer study without radiation to human subjects, and that we could simulate MC clusters with a specific range of speck sizes and analyze the dependence of the detection sensitivity on speck size. The results showed that radiologists' sensitivity in reading dnSTD images was consistently higher than reading STD images for all speck sizes. The relative gain was the largest, reaching about a factor of two, for the smallest speck size (0.125–0.150 mm) included in this study. The sensitivity was gained together with a decrease in the FP rate. The trend indicates that DL denoising is promising for increasing the differentiation between subtle MCs and noise and may potentially improve the early detection of breast cancer in DBT while reducing recalls and diagnostic workup of false detections. However, the absolute performance of the current denoiser on enhancing such subtle MCs is still low and further improvement is needed.

To use radiologists' time efficiently, we enriched each DBT volume with a large number of MC clusters, which was different from clinical DBT volumes that only occasionally contain MC clusters. In addition, the study radiologists could focus on searching for MCs in the breast phantoms without paying attention to other types of breast lesions. Both factors could affect radiologists' search task and may optimistically bias the detection sensitivity for MCs compared to reading patient DBTs in clinical settings. However, it may be expected that the relative ranking of the image conditions would be less impacted because all conditions were potentially biased in a similar way.

24734209, 2023, 10, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16439 by University Of Michigan Library, Wiley Online Library on [15/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We used the STD and STD+ modes available on a clinical DBT system to generate the low dose images and the higher dose reference condition in the observer study. The overall detection sensitivity in dnSTD was significantly higher than that in the routine STD mode and comparable to that in the STD+ mode. The VGC analysis indicated that the overall conspicuity of MCs and the radiologists' confidence ratings in the dnSTD images were significantly higher than those in the STD and STD+ images. The DL denoising approach could therefore save at least 54% of the radiation dose if the detection performance at the level of STD+ is desired. A follow up study of interest will be to evaluate if a lower dose technique than the STD mode may be used while DL denoising can still improve the MC detectability to a desired level. In addition, DBT systems from other manufacturers have been designed to operate with imaging techniques that deliver higher average glandular dose than those of the STD mode at similar breast thicknesses.[57] It is of interest to train DL denoisers to reduce noise and enhance MC detectability for the different manufacturers' systems, and to investigate whether DL denoising may enable these systems to use lower dose imaging techniques while maintaining, or improving, their image quality and performance. Reducing radiation dose to patients is an important application of denoising in medical imaging.

We applied DL denoising to the reconstructed DBT images. We did not generate synthetic mammograms from the denoised DBT volume because the synthetic mammogram technology is proprietary and vendor-dependent. How the DL denoising affects the quality of synthetic mammograms will likely depend on the imaging processing and machine vision methods used for synthesizing the 2D image. Although it is expected that the synthetic mammograms may also be benefitted from reducing noise in the DBT volume, future studies will be needed to evaluate the effects.

There are limitations in the current study. First, due to the limited availability of radiologists' time, we had to limit the variables in one observer study. We used breast phantoms of the same composition and same thickness so that the noise level varied only within a narrow range. Whether the improvement by denoising can be generalized to DBTs of a wide range of imaging properties will need to be investigated in future studies. Second, the DBT images were acquired with a single manufacturer's system and the dose levels used were selected by the automatic exposure control. Further study is needed to evaluate to what extent the improvement in MC detectability or dose reduction can be accomplished with DL denoising for different manufacturers' imaging systems. Third, the observer study was conducted using breast phantoms embedded with simulated MC clusters such that the search task of the radiologists might be different from that in clinical practice as discussed above. A future follow-up study by collecting a large set of human subject DBT containing subtle MCs will be needed to further validate the effectiveness of DL denoising for DBT.

5 | CONCLUSION

This observer study with breast phantom images demonstrates that deep-learning-based denoising has the potential to improve radiologists' sensitivity in detecting MCs in DBT and their confidence in differentiating noise from MCss without increasing the imaging dose or reading time. Future studies are needed to evaluate the generalizability of these results to the wide range of DBTs from human subjects and patient populations in clinical settings.

ACKNOWLEDGMENTS

This work is supported by National Institutes of Health award number RO1 CA214981. The authors would like to thank Prof. Magnus Båth for sharing the VGC analysis software.

CONFLICT OF INTEREST STATEMENT

The authors have no relevant conflicts of interest to disclose.

REFERENCES

1. Marinovich ML, Hunter KE, Macaskill P, Houssami N. Breast cancer screening using tomosynthesis or mammography: a meta-analysis of cancer detection and recall. *J Natl Cancer Inst*. 2018;110(9):942-949.
2. Yun SJ, Ryu C-W, Rhee SJ, Ryu JK, Oh JY. Benefit of adding digital breast tomosynthesis to digital mammography for breast cancer screening focused on cancer characteristics: a meta-analysis. *Breast Cancer Res Treat*. 2017;164(3):557-569.
3. Bahl M, Pinnamaneni N, Mercaldo S, McCarthy AM, Lehman CD. Digital 2D versus tomosynthesis screening mammography among women aged 65 and older in the United States. *Radiology*. 2019;291(3):582-590.
4. Bahl M, Mercaldo S, Dang PA, McCarthy AM, Lowry KP, Lehman CD. Breast cancer screening with digital breast tomosynthesis: are initial benefits sustained? *Radiology*. 2020;295:529-539.
5. Poplack SP, Tosteson TD, Kogel CA, Nagy HM. Digital breast tomosynthesis: initial experience in 98 women with abnormal digital screening mammography. *Am J Roentgenol*. 2007;189(3):616-623.
6. Andersson I, Ikeda DM, Zackrisson S, et al. Breast tomosynthesis and digital mammography: a comparison of breast cancer visibility and BIRADS classification in a population of cancers with subtle mammographic findings. *Eur Radiol*. 2008;18:2817-2825.
7. Spangler ML, Zuley ML, Sumkin JH, et al. Detection and classification of calcifications on digital breast tomosynthesis and 2D digital mammography: a comparison. *AJR Am J Roentgenol*. 2011;196(2):320-324.
8. Kopans D, Gavenonis S, Halpern E, Moore R. Calcifications in the breast and digital breast tomosynthesis. *Breast J*. 2011;17(6):638-644.
9. Chan H-P, Goodsitt MM, Helvie MA, et al. Digital breast tomosynthesis: observer performance of clustered microcalcification detection on breast phantom images acquired with an experimental system using variable scan angles, angular

24734209, 2023, 10, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16439 by University Of Michigan Library, Wiley Online Library on [15/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

increments, and number of projection views. *Radiology*. 2014;273(3):675-685.
10. Clauser P, Nagl G, Helbich TH, et al. Diagnostic performance of digital breast tomosynthesis with a wide scan angle compared to full-field digital mammography for the detection and characterization of microcalcifications. *Eur J Radiol*. 2016;86:2161-2168.
11. Hadjipanteli A, Elangovan P, Mackenzie A, et al. The effect of system geometry and dose on the threshold detectable calcification diameter in 2D-mammography and digital breast tomosynthesis. *Phys Med Biol*. 2017;62(3):858-877.
12. Georgian-Smith D, Obuchowski NA, Lo JY, et al. Can digital breast tomosynthesis replace full-field digital mammography? a multireader, multicase study of wide-angle tomosynthesis. *AJR Am J Roentgenol*. 2019;212(6):1393-1399.
13. Kreeger K, Smith AP, Kshirsagar A, et al. Inventors; Hologic, Inc., assignee. System and method for generating a 2D image using mammography and_or tomosynthesis image data. US patent US 10,573,276 B22020.
14. Periaswamy K, Fotin S, Inventor; iCAD, Inc., assignee. System and method for improving workflow efficiencies in reading tomosynthesis medical image data. US patent US 8,983,156 B22015.
15. van Schie G, Mann R, Imhof-Tas M, Karssemeijer N. Generating synthetic mammograms from reconstructed tomosynthesis volumes. *IEEE Trans Med Imaging*. 2013;32(12):2322-2331.
16. Wei J, Chan H-P, Helvie MA, et al. Synthesizing mammogram from digital breast tomosynthesis. *Phys Med Biol*. 2019;64(4):045011.
17. Zuckerman SP, Sprague BL, Weaver DL, Herschorn SD, Conant EF. Survey results regarding uptake and impact of synthetic digital mammography with tomosynthesis in the screening setting. *J Am College Radiol*. 2020;17(1):31-37.
18. Gilbert FJ, Tucker L, Gillan MG, et al. Accuracy of digital breast tomosynthesis for depicting breast cancer subgroups in a UK retrospective reading study (TOMMY trial). *Radiology*. 2015;277(3):697-706.
19. Lai Y-C, Ray KM, Lee AY, et al. Microcalcifications detected at screening mammography: synthetic mammography and digital breast tomosynthesis versus digital mammography. *Radiology*. 2018;289(3):630-638.
20. Bae MS, Moon WK. Is synthetic mammography comparable to digital mammography for detection of microcalcifications in screening? *Radiology*. 2018;289(3):639-640.
21. Choi JS, Han B-K, Ko EY, Kim GR, Ko ES, Park KW. Comparison of synthetic and digital mammography with digital breast tomosynthesis or alone for the detection and classification of microcalcifications. *Eur Radiol*. 2019;29(1):319-329.
22. Zuckerman SP, Conant EF, Keller BM, et al. Implementation of synthesized two-dimensional mammography in a population-based digital breast tomosynthesis screening program. *Radiology*. 2016;281(3):730-736.
23. Aujero MP, Gavenonis SC, Benjamin R, Zhang Z, Holt JS. Clinical performance of synthesized two-dimensional mammography combined with tomosynthesis in a large screening population. *Radiology*. 2017;283:70-76.
24. Caumo F, Zorzi M, Brunelli S, et al. Digital breast tomosynthesis with synthesized two-dimensional images versus full-field digital mammography for population screening: outcomes from the verona screening program. *Radiology*. 2018;287(1):37-46.
25. Zuckerman SP, Sprague BL, Weaver DL, Herschorn SD, Conant EF. Multicenter evaluation of breast cancer screening with digital breast tomosynthesis in combination with synthetic versus digital mammography. *Radiology*. 2020;297(3):545-553.
26. Regen-Tuero HC, Ram S, Gass JS, Lourenco AP. Community-based breast cancer screening using digital breast tomosynthesis versus digital mammography: comparison of screening performance and tumor characteristics. *AJR Am J Roentgenol*. 2022;218:249-257.
27. Chan H-P, Helvie MA, Klein KA, et al. Effect of dose level on radiologists' detection of microcalcifications in digital breast tomosynthesis: an observer study with breast phantoms. *Acad Radiol*. 2022;29:S42-S49.
28. Das M, Connolly C, Glick S, Gifford H. Effect of postreconstruction filter strength on microcalcification detection at different imaging doses in digital breast tomosynthesis: human and model observer studies. *Proc SPIE*. 2012;8313:831321.
29. Abdurahman S, Dennerlein F, Jerebko A, Fieselmann A, Mertelmeier T. Optimizing high resolution reconstruction in digital breast tomosynthesis using filtered back projection. *LNCS*. 2014;8539:520-527.
30. Vieira MAC, de Oliveira HCR, Nunes PF, et al. Feasibility study of dose reduction in digital breast tomosynthesis using non-local denoising algorithms. *Proc SPIE*. 2015;9412:94122C.
31. Lu Y, Chan H-P, Wei J, Hadjiiski L, Samala R. Multiscale bilateral filtering for improving image quality in digital breast tomosynthesis. *Med Phys*. 2015;42(1):182-195.
32. Liu J, Zarshenas A, Qadir A, et al. Radiation dose reduction in digital breast tomosynthesis (DBT) by means of deep-learning-based supervised image processing. *Proc SPIE*. 2018;10574:105740F.
33. Scarparo DC, Salvadeo DHP, Pedronette DCG, Barufaldi B, Maidment AD. Evaluation of denoising digital breast tomosynthesis data in both projection and image domains and a study of noise model on digital breast tomosynthesis image domain. *J Med Imaging*. 2019;6(3):031410.
34. Wolterink JM, Leiner T, Viergever MA, Isgum I. Generative adversarial networks for noise reduction in low-dose CT. *IEEE Trans Med Imaging*. 2017;36(12):2536-2545.
35. Yang Q, Yan P, Zhang Y, et al. Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE Trans Med Imaging*. 2018;37(6):1348-1357.
36. Shan H, Padole A, Homayounieh F, et al. Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. *Nat Mach Intel*. 2019;1(6):269-276.
37. Pretorius PH, Kalluri K, Liu J, et al. Multi-center evaluation of reduced dose myocardial perfusion SPECT defects employing post-filtering and deep learning denoising strategies: a human observer study. *J Nucl Med*. 2022;63(suppl 2):2400.
38. Yu Z, Rahman MA, Schindler T, et al. AI-based methods for nuclear-medicine imaging: need for objective task-specific evaluation. *J Nucl Med*. 2020;61(suppl 1):575.
39. Li K, Zhou W, Li H, Anastasio MA. Assessing the impact of deep neural network-based image denoising on binary signal detection tasks. *IEEE Trans Med Imaging*. 2021;40(9):2295-2305.
40. Samala R, Chan H-P, Hadjiiski LM, Helvie MA, Wei J, Improving detection of microcalcification clusters in low-dose digital breast tomosynthesis using deep residual learning. *Presentation at the 105th Scientific Assembly and Annual Meeting of the* Radiological Society of North America, Chicago, IL. 2019;2019:SSG13-04.
41. Sahu P, Huang H, Zhao W, Qin H, Using virtual digital breast tomosynthesis for de-noising of low-dose projection images. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI2019). 2019. doi: 10.1109/ISBI.2019.8759408:1647-1651
42. Gao M, Samala R, Fessler J, Chan H-P. Deep convolutional neural network denoising for digital breast tomosynthesis reconstruction. *Proc SPIE*. 2020;11312:113120Q.
43. Gao M, Fessler J, Chan H-P. Deep convolutional neural network regularized digital breast tomosynthesis reconstruction with detector blur and correlated noise modeling. *Proc SPIE*. 2022;12031:1203108.
44. Nakamura de Araújo D, Salvadeo DH, de Paula D. A benchmark of denoising digital breast tomosynthesis in projection

24734209, 2023, 10, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16439 by University Of Michigan Library, Wiley Online Library on [15/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

domain: neural network-based and traditional methods. *Proc SPIE*. 2022;12032:1203207.
45. de Barros Vimieiro R, Rodrigues Borges L, Barufaldi B, Maidment A, Wang G, Andrade da Costa Vieira M. Assessment of training strategies for convolutional neural network to restore low-dose digital breast tomosynthesis projections. *Proc SPIE*. 2022;12031:120311V.
46. Badal A, Cha K, Divel S, Graff C, Zeng R, Badano A. Virtual clinical trial for task-based evaluation of a deep learning synthetic mammography algorithm. *Proc SPIE*. 2019;10948:109480O.
47. Gao M, Fessler JA, Chan H-P. Deep convolutional neural network with adversarial training for denoising digital breast tomosynthesis images. *IEEE Trans Med Imaging*. 2021;40(7):1805-1816.
48. Graff CG. A new, open-source, multi-modality digital breast phantom. *Proc SPIE*. 2016;9783:978309.
49. Zhang Y, Chan H-P, Sahiner B, et al. A comparative study of limited-angle cone-beam reconstruction methods for breast tomosynthesis. *Med Phys*. 2006;33(10):3781-3795.
50. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol*. 1989;24:234-245.
51. Båth M, Månsson LG. Visual grading characteristics (VGC) analysis: a non-parametric rank-invariant statistical method for image quality evaluation. *Br J Radiol*. 2007;80(951):169-176.
52. Båth M, HJ. VGC analyzer: a software for statistical analysis of fully crossed multiple-reader multiple-case visual grading characteristics studies. *Radiat Prot Dosimetry*. 2016;169(1-4):46-53.
53. Hansson J, Månsson LG, Båth M. The validity of using ROC software for analysing visual grading characteristics data: an investigation based on the novel software VGC analyzer. *Radiat Prot Dosimetry*. 2016;169(1-4):54-59.
54. Solin LJ, Gray R, Hughes LL, et al. Surgical excision without radiation for ductal carcinoma in situ of the breast: 12-year results from the ECOG-ACRIN E5194 study. *J Clin Oncol*. 2015;33(33):3938-3944.
55. Khan S, Epstein M, Lagios MD, Silverstein MJ. Are we overtreating ductal carcinoma in situ (dcis)? *Ann Surg Oncol*. 2017;24(1):59-63.
56. Duffy SW, Dibden A, Michalopoulos D, et al. Screen detection of ductal carcinoma in situ and subsequent incidence of invasive interval breast cancers: a retrospective population-based study. *Lancet Oncol*. 2016;17(1):109-114.
57. Hendrick RE. Radiation doses and risks in breast screening. *J Breast Imag*. 2020;2(3):188-200.

How to cite this article: Chan H-P, Helvie MA, Gao M, et al. Deep learning denoising of digital breast tomosynthesis: Observer performance study of the effect on detection of microcalcifications in breast phantom images. *Med Phys*. 2023;50:6177–6189. https://doi.org/10.1002/mp.16439

24734209, 2023, 10, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16439 by University Of Michigan Library, Wiley Online Library on [15/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License