Title

Pre-whitened matched filter and convolutional neural network based model observer performance for mass lesion detection in non-contrast breast CT.

Journal
Medical Physics, 50(12)

Authors
Lyu, Su Hyun
Abbey, Craig
Hernandez, Andrew
et al.

Publication Date
2023-12-01

DOI
10.1002/mp.16685

Published in final edited form as:
Med Phys. 2023 December ; 50(12): 7558–7567. doi:10.1002/mp.16685.

Pre-whitened matched filter and convolutional neural network-based model observer performance for mass lesion detection in non-contrast breast CT

Su Hyun Lyu[1,2], **Craig K. Abbey**[3], **Andrew M. Hernandez**[2], **John M. Boone**[1,2]

[1]Department of Biomedical Engineering, University of California Davis, Davis, CA, 95618, USA.

[2]Department of Radiology, University of California Davis, Sacramento, CA, 95817, USA

[3]Department of Psychological and Brain Sciences, UC Santa Barbara, Santa Barbara, CA, 93106 USA

Abstract

Background: Mathematical model observers have been shown to reasonably predict human observer performance and are useful when human observer studies are infeasible. Recently, convolutional neural networks (CNNs) have also been used as substitutes for human observers, and studies have shown their utility as an optimal observer. In this study, a CNN model observer is compared to the pre-whitened matched filter (PWMF) model observer in detecting simulated mass lesions inserted into 253 acquired breast computed tomography (bCT) images from patients imaged at our institution.

Purpose: To compare CNN and PWMF model observers for detecting signal-known-exactly (SKE) location-known-exactly (LKE) simulated lesions in bCT images with real anatomical backgrounds, and to use these model observers collectively to optimize parameters and understand trends in performance with breast CT.

Methods: Spherical lesions with different diameters (1, 3, 5, 9 mm) were mathematically inserted into reconstructed patient bCT image data sets to mimic 3D mass lesions in the breast. 2D images were generated by extracting the center slice along the axial dimension or by slice averaging across adjacent slices to model thicker sections (0.4, 1.2, 2.0, 6.0, 12.4, 20.4 mm). The role of breast density was retrospectively studied using the range of breast densities intrinsic to the patient bCT data sets. In addition, mass lesions were mathematically inserted into Gaussian images matched to the mean and noise power spectrum of the bCT images to better understand the performance of the CNN in the context of a known ideal observer (the PWMF). The simulated Gaussian and bCT images were divided into training and testing data sets. Each training data set consisted of 91,600 images, and each testing data set consisted of 9,6000 images. A CNN and PWMF was trained on the Gaussian training images, and a different CNN and PWMF was trained

Corresponding author: John M. Boone, Ph.D., Department of Radiology, University of California Davis Health, 4860 Y Street, Ellison Building Suite 3100, Sacramento, CA 95817, (916) 735-3158, jmboone@ucdavis.edu.

Conflicts of interest:
Authors AHM, CKA, and JMB have financial interest in Izotropic Corporation, which is working on breast CT. Author JMB is editor-in-chief of Medical Physics but is blinded from the review of this paper and had no role to play in decisions regarding it.

on the bCT training images. The trained model observers were tested, and receiver operating characteristic (ROC) curve analysis was used to evaluate detection performance. The area under the ROC curve (AUC) was the primary performance metric used to compare the model observers.

Results: In the Gaussian background, the CNN performed essentially identically to the PWMF across lesion sizes and section thicknesses. In the bCT background, the CNN outperformed the PWMF across lesion size, breast density, and most section thicknesses. These findings suggest that there are higher-order features in bCT images that are harnessed by the CNN observer but are inaccessible to the PWMF.

Conclusions: The CNN performed equivalently to the ideal observer in Gaussian textures. In bCT background, the CNN captures more diagnostic information than the PWMF and may be a more pertinent observer when conducting optimal performance studies in breast CT images.

1. INTRODUCTION

Breast computed tomography (bCT) is a relatively new breast imaging modality based upon cone-beam CT geometry[1,2], or helical CT geometry [3]. While a number of human observer studies have been published on the performance of breast CT [1,4,5], such studies can be limited when assessing a large number of images, which is often the case when fine-tuning an imaging system or identifying optimal parameter settings for lesion detectability. Mathematical model observers have been shown to reasonably predict human observer performance [6,7], and are useful when human observer studies are infeasible [8]. Recently, convolutional neural networks (CNNs) have also been used as substitutes for human observers [9, 10], where they are referred to as anthropomorphic models. Studies have shown that appropriately trained CNNs have utility as an optimal observer, the so-called ideal observer [11–13]. Ideal observers are useful for assessing how much diagnostic information is contained in an image in advance of processing or display effects that make this information accessible to human observers [14–16].

In this study, a CNN model observer is compared to a more conventional ideal-observer model, the pre-whitened matched filter (PWMF). The PWMF has an appealing definition that involves the signal to be detected as well as the texture of the image background, as specified by the power spectrum. The PWMF is known to be an optimal detection filter for images with variability that is described by a stationary Gaussian distribution [14,17,18] and it is related to image-quality measures like noise-equivalent quanta and detective quantum efficiency. However, it has been demonstrated that breast CT images are not Gaussian distributed [19], and so it is not clear that the PWMF represents an ideal observer in this case. This means that it may be possible that the PWMF is systematically missing diagnostic information in breast CT images, and may therefore underestimate optimal performance across patient and imaging factors.

Neural networks [20] have been suggested as a way to implement the ideal observer when the analytical approach of defining a likelihood ratio is not feasible, and more recently CNNs specifically have been evaluated for this purpose [21,22]. This general approach is based on the flexibility of the network architecture and the ability to train the model on samples of data rather than derive accurate probabilistic models of images with complex non-Gaussian statistical properties. Flexible network architectures allow the network models to extract image information that may not be accessible to a model like the PWMF that is constrained to be linear.

The field has generally found that network models are capable of higher performance than an optimal linear filter like the PWMF, and this includes recent studies by Baek and colleagues [23–25] using synthetic bCT images with simulated anatomical backgrounds. In this report, we seek to extend these results to bCT images acquired from patients, for the task of detecting a simulated lesion that is embedded in real anatomical background. This, to our knowledge, is the first comparison between CNNs and mathematical model observers in bCT images with real anatomical backgrounds, and may provide a more accurate assessment of the model observers when applied to bCT. This detection paradigm has been studied previously using the PWMF [26], and was found useful for understanding how detection performance is dependent on the interaction between lesion diameter and section thickness.

This study involves a data set of 322 patient breast CT images acquired at the UC Davis Medical Center under an IRB approved protocol [1,5]. We also investigate Gaussian images matched to the mean and power-spectrum of the breast CT images. This allows us to implement our particular CNN in an imaging condition where the PWMF is known to be an ideal observer. If the CNN is able to closely approximate the ideal observer, then we have some confidence that the architecture and training process is adequate for comparing performance more generally. In the breast CT images, we compare the CNN to the PWMF across lesion diameter, slice thickness, and breast-density categories.

2. METHODS

2.1. Image generation

2.1.1 Insertion of lesions into breast background—

Spherical lesions were mathematically inserted into reconstructed patient bCT image volumes to mimic 3D mass lesions in breast parenchyma. Human bCT images were acquired at the UC Davis Medical Center under IRB-approved clinical trials which recruited patients receiving BIRADS 4 or 5 on their breast screening exams [1,5]. Enrolled patients were scanned on prototype bCT scanners [27] developed in our laboratory. All patients subsequently underwent biopsy to yield the ground truth diagnosis for suspicious lesions as benign or malignant. Of the four existing iterations of prototype bCT scanners, the first two scanners, which are very similar in design, were used to scan a total of 322 women. From this cohort we selected 253 image volume data sets for this study on the basis of not containing artifacts and not involving contrast-imaging. Each volume data set contained 300–500 reconstructed slices (512 × 512 matrix size) with isotropic voxel sizes of 0.4 mm.

A previously published method developed by Packard *et al.* [26] was used to insert spherical lesions into bCT images and is briefly detailed here. Let $f[i, j, k]$ be a reconstructed bCT image volume and $s[i, j, k]$ be a segmented version of $f[i, j, k]$, where each voxel is segmented as adipose tissue, fibroglandular tissue, skin, or air [28]. $s[i, j, k]$ is used to compute $TB[i, j, k]$, a binary volume identifying tissue boundaries (TB) in the breast, and having a value of 1 for every voxel segmented as glandular tissue but having one of its six adjacent voxels (in 3D) segmented as adipose tissue, and 0 for all other voxels. $TB[i, j, k]$ is used to compute $d_{TB}[i, j, k]$, the distance from every voxel to the nearest tissue boundary identified in $TB[i, j, k]$. $d_{TB}[i, j, k]$ is signed such that it is positive for voxels segmented as adipose tissue and negative for voxels segmented as glandular tissue.

Let the index location $[i_{LC}, j_{LC}, k_{LC}]$ be a randomly generated lesion center "LC" where the lesion is to be inserted. The lesion location $[i_{LC}, j_{LC}, k_{LC}]$ is kept if the surrounding 64 × 64 × 64 volume is fully contained within the patient breast and does not contain skin; otherwise, the lesion center coordinates are re-generated. Let D be the diameter of the lesion to be inserted and $d_{LC}[i, j, k]$ be the distance from each voxel to lesion center $[i_{LC}, j_{LC}, k_{LC}]$. The distance to the nearest lesion boundary $d_{LB}[i, j, k]$ is then defined as:

$$d_{LB}[i, j, k] = \frac{1}{2}D - d_{LC}[i, j, k] \tag{1}$$

and is positive for voxels within the spherical lesion and negative for voxels outside the spherical lesion.

Let $\Delta\ I$ be the mean differential intensity between all glandular and adipose voxels in the image. Then, the resulting image volume with the inserted spherical lesion $f_{sim}[i, j, k]$ is:

$$f_{sim}[i, j, k] = f[i, j, k] + (\ \Delta\ I \times M(d_{TB}[i, j, k]) \times M(d_{LB}[i, j, k])) \tag{2}$$

Intensity is added on a voxel-by-voxel basis in order to preserve the native image noise. Outside of the spherical lesion, the added term becomes zero. The added intensity at each voxel is modulated by $M(d_{TB}[i, j, k])$, the tissue-boundary modulation term, which ranges from 0–1 and approaches zero when $d_{TB}[i, j, k]$ is negative. In effect, this term allows intensity only to be added to adipose regions and smooths the regions within the inserted lesion where adipose and glandular tissue coincide. The added intensity at each voxel is further modulated by $M\left(\frac{1}{2}D - d_{LC}[i, j, k]\right)$, the lesion-boundary modulation term, which also ranges from 0–1 and serves to smooth the edge of the spherical inserted lesion. The modulation function M is derived by mathematically modeling the edge-blurring at boundaries between adipose and glandular tissue in the native patient image. These modulation terms serve to retain the native image resolution (~ modulation transfer function).

A 2D 64 × 64 × 1 image was then generated from the 3D image volume by extracting the center slice along the axial dimension or by slice averaging across adjacent slices to model thicker sections. Previous model observer studies demonstrated that higher detection performance was found in the axial and sagittal views compared to the coronal view in bCT [26]. The resulting patch with the added lesion is denoted as $I_n^+(x, y)$, where n represents the

n^{th} lesion, and the same patch without the added signal is denoted as $I_n^-(x, y)$. Four lesion diameters (1, 3, 5, 9 mm) and six section thicknesses (0.4, 1.2, 2.0, 6.0, 12.4, 20.4 mm) were studied. Sample lesion-present patches with varying lesion diameters and section thicknesses are shown in Figure 1.

The role of breast density was retrospectively studied using the range of breast densities spanning the patient bCT data sets. For every patient, breast density was quantified by the volumetric glandular fraction (VGF). Let n_g represent the number of voxels segmented as glandular tissue and n_a represent the number of voxels segmented as adipose tissue in the segmentation volume $s[i, j, k]$. The VGF is then defined as:

$$VGF = \frac{n_g}{n_g + n_a} \tag{3}$$

2.1.2. Insertion of lesions into Gaussian background—

Mass lesions were mathematically inserted into Gaussian images matched to the mean and noise power spectrum of the bCT images. For bCT images a 3D lesion was inserted into a 3D background volume. In comparison, for Gaussian images a 2D lesion (derived from the bCT patches) was inserted into a 2D background patch. Let $\overline{I^+}(x, y)$ and $\overline{I^-}(x, y)$ denote the mean lesion-present and mean lesion-absent patches, respectively. The added lesion is the mean signal $\overline{S}(x, y)$ across image patches from all bCT training images:

$$\overline{S}(x, y) = \overline{I^+}(x, y) - \overline{I^-}(x, y) \tag{4}$$

This mean signal inherently smooths the lesion boundary and dampens the signal in thicker sections.

Generating the 2D Gaussian background patch requires knowledge of the bCT power spectrum. The power spectrum was estimated from the training bCT background images. Let $I_n^-[x, y]$ represent the n^{th} lesion-absent patch and $H[x, y]$ represent a 2D Hamming filter. The windowed deviation function $\Delta\, I_n^-(x, y)$ is then defined as:

$$\Delta\, I_n^-(x, y) = H[x, y] \times (I_n^-[x, y] - \overline{I^-}[x, y]) \tag{5}$$

where $H[x, y]$ is a windowing function used to attenuate artifacts arising from the cyclic nature of the discrete Fourier transform as it approaches the edge of an image [7]. Let N represent the total number of lesion-absent patches. The mean power spectrum $\overline{PS}(f_x, f_y)$ of the image backgrounds is then estimated as:

$$\overline{PS}(f_x, f_y) = \frac{1}{N - 1}\sum_{n=1}^{N}\left|\Delta\, \widehat{I_n^-}(x, y)\right|^2 \tag{6}$$

where the caret is used to represent the Fourier transform. The mean power spectrum was then normalized to compensate for the windowing that occurs in Equation 5. To normalize the power spectrum, the mean pixel variance at each pixel across all the background images was first computed:

$$V = \frac{1}{N-1}\sum_{n=1}^{N}\left(I_n^-[x, y] - \overline{I^-}[x, y]\right)^2 \tag{7}$$

Then, the normalized mean noise power spectrum $\overline{PS}(f_x, f_y)_{norm}$ is defined as:

$$\overline{PS}(f_x, f_y)_{norm} = \frac{\sum_{x,y} V}{\sum_{x,y}\overline{PS}(f_x, f_y)} \times \overline{PS}(f_x, f_y) \tag{8}$$

$\overline{PS}(f_x, f_y)_{norm}$ was then converted to the spatial domain using a 2D inverse Fourier transform. Gaussian background patches were then generated by convolving the square root of $\overline{PS}(x, y)_{norm}$ with a 64 × 64× 1 patch of random white noise $E(x, y)$. In using the normalized mean noise power spectrum of bCT images for Gaussian simulation, pixel variance between the two image types is maintained. The final Gaussian patch $G^+\left(x, y\right)$ with the inserted lesion is defined as:

$$G^+(x, y) = \overline{S}(x, y) + \left(E(x, y) * \sqrt{\overline{PS}(x, y)_{norm}}\right) \tag{9}$$

and the same patch without the inserted lesion is defined as $G^-(x, y)$. Sample Gaussian lesion-present patches with varying lesion diameters and section thicknesses are shown in Figure 1.

2.2. Model observers: Pre-whitened matched filter (PWMF)

2.2.1. PWMF computation—The PWMF is a mathematical model observer that makes use of the mean signal and background power spectrum of a set of images to compute a decision variable [16], [17]. A unique filter was computed for each combination of lesion diameter, section thickness, and background condition (bCT or Gaussian) to tune the filter to the environment. Let FT denote the 2D Fourier transform, and FT^{-1} denote the 2D inverse Fourier transform. Let $\overline{S}(x, y)$ be the mean signal across all *training* images, and $\overline{PS}(f_x, f_y)_{norm}$

be the normalized mean noise power spectrum of *training* image backgrounds. The PWMF $w[x, y]$ is then defined as:

$$w[x, y] = FT^{-1}\left\{\frac{FT\{\overline{S}(x, y)\}}{\overline{PS}(f_x, f_y)_{norm}}\right\} \tag{10}$$

Once the PWMF was computed from a set of *training* images, it was then applied to an independent set of *testing* image patches in order to evaluate lesion detection performance. For bCT conditions, lesion-present and lesion-absent testing patches were generated from completely separate bCT volume data sets using the lesion insertion process described in Section 2.1. Let $I_n[x, y]$ represent a testing image patch and $w[x, y]$ be the PWMF tuned to that specific lesion diameter, section thickness, and background condition. A scalar-valued decision variable λ_n was then computed for each testing patch as follows:

$$\lambda_n = \sum_{x, y} I_n[x, y] \times w[x, y] \tag{11}$$

2.2.2. Training and testing the PWMF—Out of the 253 total bCT volume data sets, 229 data sets ($N = 229$, ~90%) were used for training the PWMF. For each training data set, 200 unique lesion centers were identified and used to generate 200 lesion-present patches and 200 lesion-absent patches. The remaining 24 data sets ($K = 24s$, ~10%) were used for testing the PWMF. For testing, 200 unique lesion centers were first identified to generate 200 lesion-present patches, and 200 additional different lesion centers were identified to generate 200 lesion-absent patches so as not to correlate the decision variables. In total, for a given lesion diameter and section thickness, 91,600 bCT training patches (i.e., 229 × 400) were generated and 9,600 bCT testing patches (i.e., 24 × 400) were generated. Though simulated Gaussian image patches were not dependent on patient data sets or lesion centers, the same number of training and testing data sets were generated to match the simulated bCT image data set.

2.3. Model observers: Convolutional neural network (CNN)

2.3.1 CNN architecture: A convolutional neural network (CNN) was implemented to perform a simple binary classification task and compute a decision variable. The input to the CNN was a single 2D image patch, and the output was a scalar-valued decision variable between 0 and 1, scaled by the sigmoid function. The network consisted of two convolutional layers followed by one fully connected layer. The two convolutional layers served to extract feature maps from the preceding layer, and the fully connected layer condensed the feature maps into a scalar-valued decision variable. The first convolutional layer contained 3 × 3 filters with a stride of 1, and the second convolutional layer contained 3 × 3 filters with a stride of 1. Batch normalization was implemented after the first convolutional layer. Max pooling layers were implemented after each convolutional layer with a pool size of 2 × 2. Dropout was implemented after the first max pooling layer with a rate of 0.2, after the second max pooling layer with a rate of 0.2, and after the fully connected layer with a rate of 0.5. The rectified linear unit (ReLU) activation function was used in all layers, including the fully connected layer.

The choice of a three-layered architecture was due to the relatively simple task of binary classification in a signal-known-exactly (SKE) setting. Increasing the depth of the network could have resulted in a more complex model but may have led to overfitting [29]. In this model, the total number of parameters was 821,889. A diagram of the CNN architecture is shown in Figure 2.

2.3.2 Training and testing the CNN: The same training and testing simulated data sets described in Section 2.2.2 for bCT and Gaussian background images were used to train and test the CNN model observer. For each combination of lesion diameter, section thickness, and background condition, 91,600 images were used for training and validation, and 9,600 images were used for testing.

The CNN was trained to minimize the binary-cross entropy (BCE) loss. Let y be the ground truth label (0 or 1), $\widetilde{y}$ be the predicted value, and N be the number of samples. BCE loss is then defined as:

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\widetilde{y}) + (1 - y_i)\log(1 - \widetilde{y_i})\right] \tag{12}$$

The training metric was accuracy:

$$Accuracy = \frac{\#\ correct\ predictions}{\#\ total\ predictions} \tag{13}$$

The Adam optimizer [30] was used with a learning rate of 1e-5 and a batch size of 64. The maximum number of training epochs was set to 150 but early stopping was implemented such that if the validation loss did not decrease after 4 epochs, training was stopped. The CNN model was implemented in Python using the Keras library [31]. An NVIDIA GeForce GTX 1080 GPU was used.

2.4. Performance evaluation and statistical analysis—For detection performance evaluation, receiver operating characteristic (ROC) curve analysis was used on the scalar-valued decision variables produced by the model observers. For a range of thresholds that discriminated each variable as true positive, true negative, false positive, or false negative, the true positive rate was plotted against the false positive rate to produce an empirical ROC curve. The area under the ROC curve (AUC) was computed individually for each testing data set ($\mathrm{K} = 24$) instead of from one pool of all decision variables in order to be able to study the effect of breast density on individual bCT images. In addition, the mean and standard deviation of AUCs across all 24 data sets were also computed to characterize the average detectability for that combination of parameters. Mean AUCs were plotted with 95% confidence error bars in Section 3. Let $\overline{AUC}$ and σ be the mean and standard deviation, respectively, of AUCs across the K testing data sets. The 95% confidence interval CI_{95} is then defined as:

$$CI_{95} = \overline{AUC} \pm 1.96 \times \frac{\sigma}{\sqrt{K}} \tag{14}$$

In Section 3.1, PWMF and CNN detection performance was compared on Gaussian background images. To quantify the similarity between the model observers, the maximum absolute difference between individual and mean AUCs were computed across clinical parameters. The maximum absolute difference between individual PWMF and CNN AUCs from 24 testing data sets is defined as:

$$|\Delta AUC|_{max} = \max\left(|AUC_{PWMF} - AUC_{CNN}|\right) \tag{15}$$

In Section 3.2, PWMF and CNN detection performance was compared across clinical parameters using paired t-tests. To address the multiple comparisons problem, we employed the Bonferroni correction to adjust the family-wise error rate. One asterisk (*) is used to indicate $p < .05$ and two asterisks (**) are used to indicate $p < .01$.

3. RESULTS

3.1 Comparison of PWMF and CNN model observers in Gaussian background

The CNN observer closely matched the PWMF observer in detection performance across all section thicknesses and lesion diameters. Across these parameters, $|\Delta AUC|_{max}$ was 0.0096. PWMF and CNN detection performance on Gaussian images for the native section thickness (0.4 mm) as a function of lesion diameter are displayed in Figure 3a. Detection performance is nearly identical. Figures 3b–c show PWMF and CNN detection performance on Gaussian images as a function of section thickness for 1- and 5-mm lesions. In these settings, the CNN observer also closely aligned with the PWMF observer.

3.2 Comparison of PWMF and CNN model observers in bCT background

3.2.1 Model observer comparison across lesion diameter—

Model observer performance on bCT images displayed in the native section thickness ($Z = 0.4$ mm) were averaged across breast densities and plotted as a function of lesion diameter in Figure 4a. Across all lesion diameters, the CNN outperformed the PWMF ($p < .01$). These findings were further analyzed in the context of breast density in Figures 4b–c. Of the 24 testing data sets, bCT patches extracted from patients with lower VGF breasts ($N = 12$) were evaluated and the mean AUC was plotted in Figure 4b, and bCT patches extracted from patients with higher VGF breasts $(\mathrm{N} = 12)$ were evaluated and the mean AUC was plotted in Figure 4c. The range of breast densities in the testing data set was [0.118, 0.597]. The CNN consistently outperformed the PWMF in higher- and lower-density breasts $(p < .05)$.

3.2.2 Model observer comparison across section thickness—

Model observer performance on bCT images containing 1- and 5- mm lesions were averaged across all breast densities and plotted as a function of section thickness in Figures 5a and 6a, respectively. Across all section thicknesses for both lesion sizes, the CNN outperformed the PWMF $(p < .01)$. These findings were further analyzed in the context of breast density in Figures 5b–c and 6b–c. Of the 24 testing data sets, bCT patches extracted from patients with lower VGF breasts $(\mathrm{N} = 12)$ were evaluated and the mean AUC was plotted in Figures 5b and 6b, and bCT patches extracted from patients with higher VGF breasts $(\mathrm{N} = 12)$ were evaluated and the mean AUC was plotted in Figure 5c and 6c. The CNN outperformed the PWMF in higher- and lower-density breasts for the detection of both lesion sizes $(p < .05)$ except when detecting the 1 mm in high VGF breasts in section thicknesses of 1.2 and 2 mm. As might be expected, both observers detected larger lesions better than smaller lesions.

Figure 5 shows that the thinnest section (0.4 mm) is not the ideal display thickness for detecting small (≤ 1 mm) lesions. Rather, the 1.2- and 2-mm section thicknesses enabled peak detection for both model observers. For the 5 mm lesion (Figure 6), detectability was minimally affected by section thickness, and detection performance decreased only slightly when section thickness exceeded 6 mm. In Figure 6, we observe that CNN detection performance of 5 mm lesions is minimally dependent on breast density. In comparison, the PWMF detection performance decreases for higher VGF.

4. DISCUSSION

In this study, a CNN model observer was compared to a more conventional model observer, the PWMF, for the binary detection task of detecting a SKE mass lesion. The PWMF is a well-established linear detection filter that has proven to be the ideal observer in background conditions with stationary Gaussian noise. CNNs have likewise been implemented as potential ideal observers given their ability to learn and extract relevant features from images even in the presence of complex backgrounds.

The model observers were used to detect mass lesions in Gaussian background to better understand the performance of the CNN in the context of a known ideal observer (the PWMF). The added signal and simulated Gaussian noise were matched to the mean signal and mean noise power spectrum of bCT images. We observed that the CNN performed nearly identically to the PWMF across lesion sizes and section thicknesses, with the largest absolute difference between AUCs being 0.0096. This result suggests that the CNN is effectively an ideal observer in Gaussian textures and that it extracts at least as much diagnostic information from an image as the PWMF does. It is likely that the CNN model can be applied to non-Gaussian imaging conditions and at minimum it would perform at least as well as the PWMF.

The model observers were used to detect mass lesions mathematically inserted into bCT images with real anatomical background. The CNN outperformed the PWMF in detecting mass lesions across lesion size, most section thicknesses, and breast density. As expected, both observers detected larger lesions better than smaller lesions. For both observers, the optimal section thickness of display was between 1.2 – 2.0 mm, the equivalent of 3–5 reconstructed slices, and reduced detection performance was observed in thinner and thicker sections. We suspect that quantum noise interference contributes to reduced detection performance in thinner sections, and that anatomical noise contributes to reduced detection performance in thicker sections due to the superposition of glandular anatomy. Packard et al. observed similar results in a previous study [26].

It is notable that the CNN outperformed the PWMF in bCT background while in Gaussian conditions the two observers performed equally. These findings suggest that there is a substantial amount of diagnostic information in bCT images that the CNN captures that is not accessible to the PWMF. The PWMF employs only the mean signal and mean noise power spectrum of an image to formulate a decision variable. The PWMF and linear model observers in general are evidently limited. In contrast, neural networks, which are non-linear, can perform (perhaps ideally) in an SKE setting. Our results should serve as motivation for future studies that identify the specific informative features that allow the CNN to outperform the PWMF (e.g., using reverse-correlation methods [32]).

Previous studies have demonstrated that the PWMF is the ideal observer in Gaussian image backgrounds [14,17,18]. This study suggests that in non-Gaussian backgrounds (such as bCT), the PWMF fails to recognize higher order statistical information and image features, whereas the CNN clearly yields superior performance. Hence, a CNN observer may be more appropriate when estimating peak performance across patient and imaging factors.

This study had limitations. Model observers were compared for the detection of relatively simple signals: SKE-LKE mass lesions. Microcalcifications were not simulated in this study, and an appropriate simulation would require more complex modeling of partial volume effects. Previously, human observer studies have indicated that microcalcifications are more difficult to detect than mass lesions in breast CT [1]. Therefore, to fully understand the utility of PWMF and CNN model observers in breast CT, an evaluation of their abilities to detect microcalcifications is necessary. Future studies will investigate this. A three-layered CNN was used in this study. We recognize that for more complex detection tasks, a deeper architecture may be required and that additional training methods such as transfer learning may be useful. Furthermore, we recognize that the CNN performed exceptionally well in this study primarily due to two reasons: 1) the signal was known exactly, and 2) there was a large amount of labeled training data. These conditions were only possible since we mathematically simulated the training signals, albeit on actual breast anatomy. Finally, we chose the PWMF among other linear observers for this study. Future studies comparing CNNs with other linear observers such as the Hotelling observers may be useful to underscore the utility of the CNN-based observer in bCT.

CONCLUSION

In this study, we used PWMF and CNN model observers to detect SKE mass lesions in patient bCT images. The CNN outperformed the PWMF across lesion size, most section thicknesses, and breast density. We conclude that the CNN captures more diagnostic information from bCT images than the PWMF and may be a more suitable observer when conducting optimal performance studies.

While model observer studies are important, they do not fundamentally replace the need for human observer studies. However, there is an increasing emphasis on virtual clinical trials in the literature. The power in these model observer studies is of course that many more lesions and lesion placements can be studied than with human observers, and this provides the ability to generate statistically meaningful results which can aid in optimizing breast CT parameters prior to human observer studies.

Acknowledgements:

This study was funded in part from the NIH: R01 EB025829 (UCSB), R01 CA214515 (UNM), AWD101462-q (MIDRC), and R01 CA181081 (UCD).

REFERENCES

1. Lindfors KK, Boone JM, Nelson TR, Yang K, Kwan ALC, Miller DF. Dedicated Breast CT: Initial Clinical Experience. Radiology. 2008;246(3):725–733. doi:10.1148/radiol.2463070410 [PubMed: 18195383]
2. O'Connell A, Conover DL, Zhang Y, et al. Cone-Beam CT for Breast Imaging: Radiation Dose, Breast Coverage, and Image Quality. Am J Roentgenol. 2010;195(2):496–509. doi:10.2214/AJR.08.1017 [PubMed: 20651210]
3. Kalender WA, Beister M, Boone JM, Kolditz D, Vollmar SV, Weigel MCC. High-resolution spiral CT of the breast at very low dose: concept and feasibility considerations. Eur Radiol. 2012;22(1):1–8. doi:10.1007/s00330-011-2169-4 [PubMed: 21656331]
4. Chen L, Boone JM, Abbey CK, et al. Simulated lesion, human observer performance comparison between thin-section dedicated breast CT images versus computed thick-section simulated projection images of the breast. Phys Med Biol. 2015;60(8):3347–3358. doi:10.1088/0031-9155/60/8/3347 [PubMed: 25825980]
5. Prionas ND, Lindfors KK, Ray S, et al. Contrast-enhanced Dedicated Breast CT: Initial Clinical Experience. Radiology. 2010;256(3):714–723. doi:10.1148/radiol.10092311 [PubMed: 20720067]
6. Eckstein MP, Abbey CK, Whiting JS. Human vs model observers in anatomic backgrounds. In: Kundel HL, ed. ; 1998:16–26. doi:10.1117/12.306180
7. Burgess AE, Jacobson FL, Judy PF. Human observer detection experiments with mammograms and power-law noise. Med Phys. 2001;28(4):419–437. doi:10.1118/1.1355308 [PubMed: 11339738]
8. Hernandez AM, Becker AE, Hyun Lyu S, Abbey CK, Boone JM. High-resolution μCT imaging for characterizing microcalcification detection performance in breast CT. J Med Imaging. 2021;8(05). doi:10.1117/1.JMI.8.5.052107
9. Fan F, Ahn S, Man BD, et al. Deep learning-based model observers that replicate human observers for PET imaging. In: Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment. Vol 11316. SPIE; 2020:53–58. doi:10.1117/12.2547505
10. Brankov JG, Yongyi Yang, Liyang Wei, El Naqa I, Wernick MN. Learning a Channelized Observer for Image Quality Assessment. IEEE Trans Med Imaging. 2009;28(7):991–999. doi:10.1109/TMI.2008.2008956 [PubMed: 19211351]
11. Barlow HB. Increment thresholds at low intensities considered as signal/noise discriminations. J Physiol. 1957;136(3):469–488. doi:10.1113/jphysiol.1957.sp005774 [PubMed: 13429514]

12. Burgess AE, Wagner RF, Jennings RJ, Barlow HB. Efficiency of Human Visual Signal Discrimination. Science. 1981;214(4516):93–94. doi:10.1126/science.7280685 [PubMed: 7280685]

13. Burgess A Image quality, the ideal observer, and human performance of radiologic decision tasks. Acad Radiol. 1995;2(6):522–526. doi:10.1016/S1076-6332(05)80411-8 [PubMed: 9419600]

14. Wagner RF, Brown DG. Unified SNR analysis of medical imaging systems. Phys Med Biol. 1985;30(6):489–518. doi:10.1088/0031-9155/30/6/001 [PubMed: 29081545]

15. Vennart W ICRU Report 54: Medical imaging-the assessment of image quality - ISBN 0–913394-53-X. April 1996, Maryland, U.S.A. Radiography. 1997;3(3):243–244. doi:10.1016/S1078-8174(97)90038-9

16. Barrett HH, Abbey CK, Clarkson E. Objective assessment of image quality III ROC metrics, ideal observers, and likelihood-generating functions. J Opt Soc Am A. 1998;15(6):1520. doi:10.1364/JOSAA.15.001520

17. Sharp P, Barber DC, Brown DG, et al. Appendix E: The Hotelling Observer. Rep Int Comm Radiat Units Meas. 1996;os-28(1):63–65. doi:10.1093/jicru_os28.1.63

18. Barrett HH, Yao J, ROLLANDt JP, MYERSt KJ. Model observers for assessment of image quality. Proc Natl Acad Sci USA. Published online 1993.

19. Abbey CK, Nosratieh A, Sohl-Dickstein J, Yang K, Boone JM. Non-Gaussian statistical properties of breast images: Non-Gaussian statistical properties of breast images. Med Phys. 2012;39(11):7121–7130. doi:10.1118/1.4761869 [PubMed: 23127103]

20. Boone JM, Gross GW, Greco-Hunt V. Neural Networks in Radiologic Diagnosis. Invest Radiol. 1990;25(9):1012–1016. [PubMed: 2211042]

21. Kupinski MA, Edwards DC, Giger ML, Metz CE. Ideal observer approximation using Bayesian classification neural networks. IEEE Trans Med Imaging. 2001;20(9):886–899. doi:10.1109/42.952727 [PubMed: 11585206]

22. Zhou W, Li H, Anastasio MA. Approximating the Ideal Observer and Hotelling Observer for binary signal detection tasks by use of supervised learning methods. IEEE Trans Med Imaging. 2019;38(10):2456–2468. doi:10.1109/TMI.2019.2911211 [PubMed: 30990425]

23. Han M, Baek J. A convolutional neural network-based anthropomorphic model observer for signal-known-statistically and background-known-statistically detection tasks. Phys Med Biol. 2020;65(22):225025. doi:10.1088/1361-6560/abbf9d [PubMed: 33032268]

24. Kim B, Han M, Baek J. A Convolutional Neural Network-Based Anthropomorphic Model Observer for Signal Detection in Breast CT Images Without Human-Labeled Data. IEEE Access. 2020;8:162122–162131. doi:10.1109/ACCESS.2020.3021125

25. Kim G, Han M, Shim H, Baek J. A convolutional neural network-based model observer for breast CT images. Med Phys. 2020;47(4):1619–1632. doi:10.1002/mp.14072 [PubMed: 32017147]

26. Packard NJ, Abbey CK, Yang K, Boone JM. Effect of slice thickness on detectability in breast CT using a prewhitened matched filter and simulated mass lesions: Effect of slice thickness on breast CT detectability. Med Phys. 2012;39(4):1818–1830. doi:10.1118/1.3692176 [PubMed: 22482604]

27. Gazi PM, Yang K, Burkett GW, Aminololama-Shakeri S, Anthony Seibert J, Boone JM. Evolution of spatial resolution in breast CT at UC Davis: Evolution of spatial resolution in breast CT at UC Davis. Med Phys. 2015;42(4):1973–1981. doi:10.1118/1.4915079 [PubMed: 25832088]

28. Ghazi P, Hernandez AM, Abbey C, Yang K, Boone JM. Shading artifact correction in breast CT using an interleaved deep learning segmentation and maximum-likelihood polynomial fitting approach. Med Phys. 2019;46(8):3414–3430. doi:10.1002/mp.13599 [PubMed: 31102462]

29. Jimmy L, Caruana R. Do Deep Nets Really Need to be Deep?

30. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. Published online January 29, 2017. Accessed January 26, 2023. http://arxiv.org/abs/1412.6980

31. Chollet F. Keras. Published online 2015. https://github.com/fchollet/keras

32. Abbey CK, Sengupta S, Zhou W, et al. Analyzing neural networks applied to an anatomical simulation of the breast. In: Mello-Thoms CR, Taylor-Phillips S, eds. Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment. SPIE; 2022:14. doi:10.1117/12.2612614

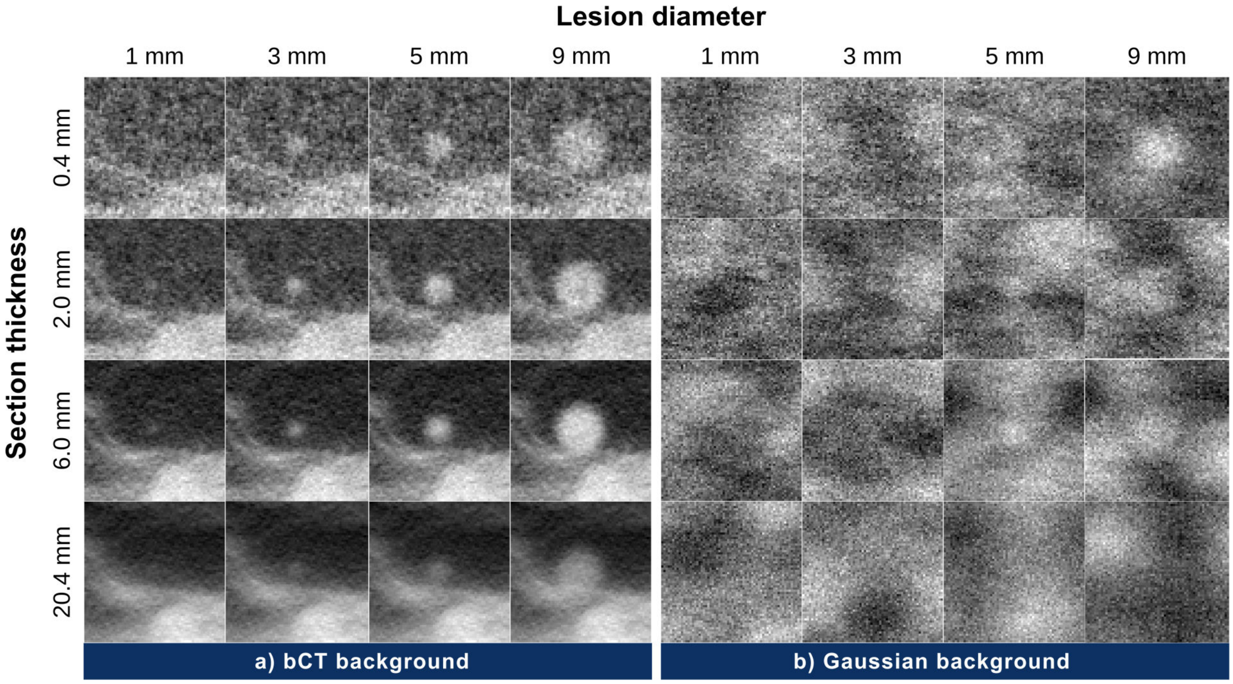

Figure 1.

Example lesion-present patches in a) patient bCT background and b) simulated Gaussian background for varying lesion diameters and section thicknesses.

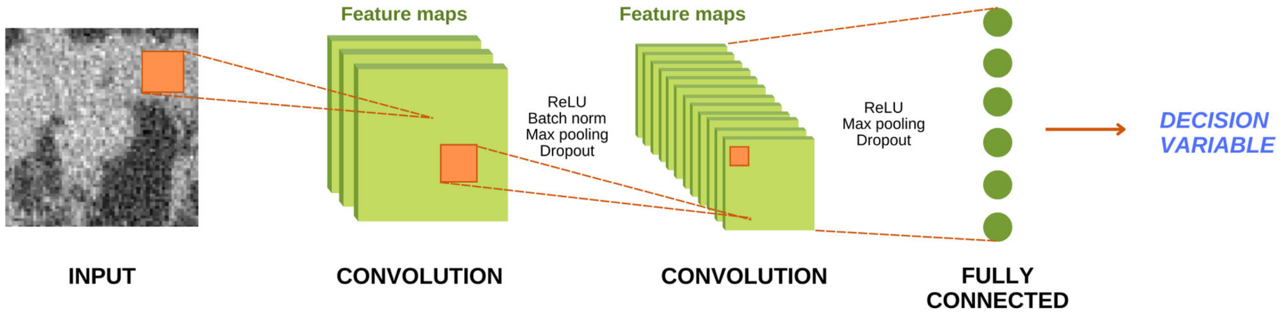

Figure 2.

Convolutional neural network architecture.

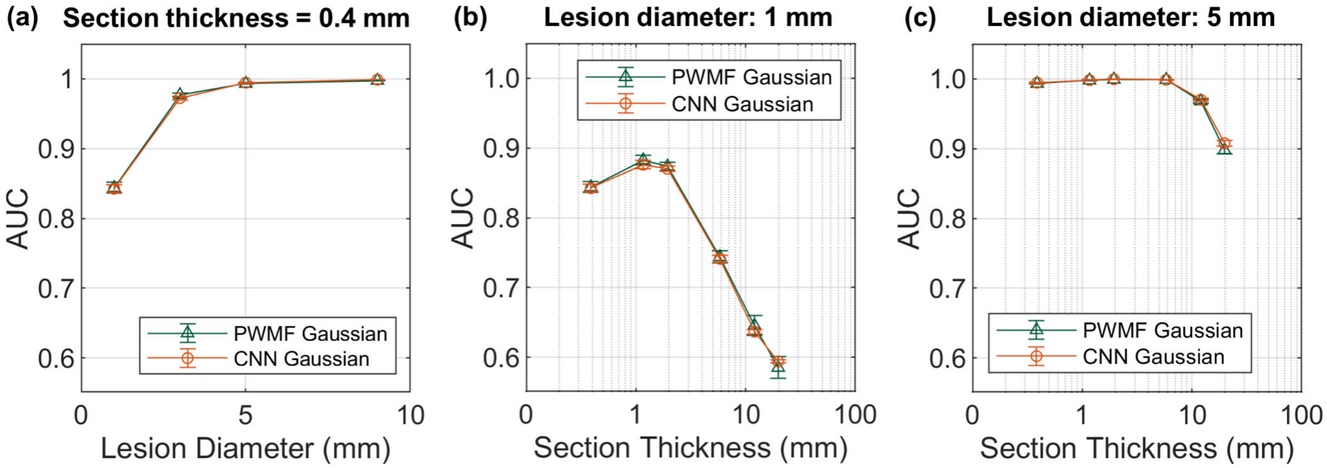

Figure 3.

Comparison of PWMF and CNN model observers on Gaussian background images as a function of **(a)** lesion diameter, displayed in the native section thickness (0.4 mm), **(b)** section thickness for a 1-mm lesion, and **(c)** section thickness for a 5-mm lesion. Detection performance is nearly identical across all parameters. Error bars correspond to 95% confidence intervals for each performance estimate.

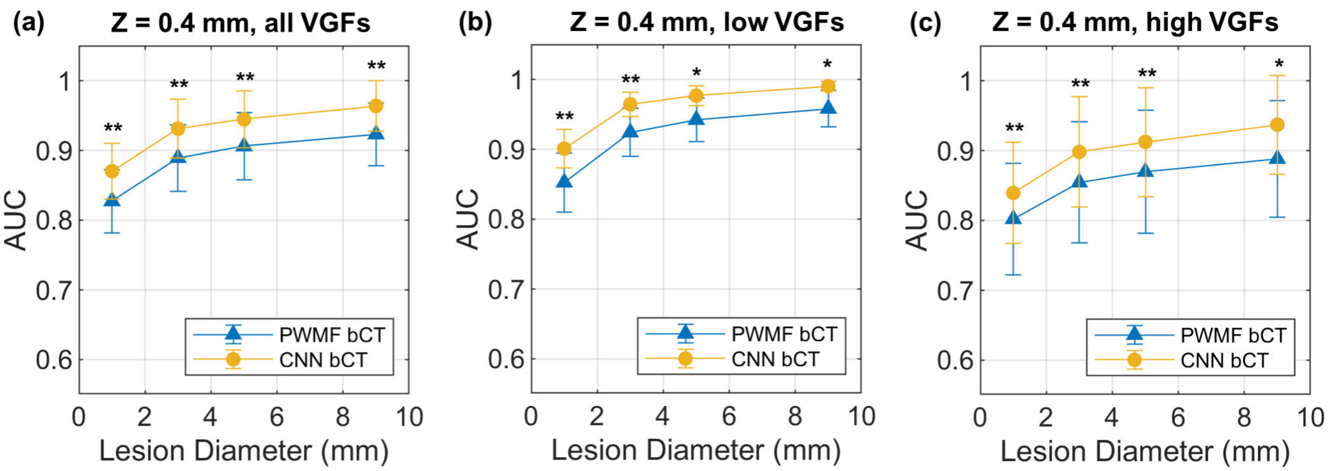

Figure 4.

Comparison of PWMF and CNN model observers on bCT images displayed in the native section thickness $(Z = 0.4\text{ mm})$ as a function of lesion diameter across **(a)** all VGFs $(N = 24)$, **(b)** low VGFs $(N = 12)$, and **(c)** high VGFs $(N = 12)$. Paired t-tests were used with Bonferroni correction to adjust for multiple comparisons. One asterisk (*) is used to indicate $p < .05$, and two asterisks (**) are used to indicate $p < .01$. Error bars correspond to 95% confidence intervals for each performance estimate.

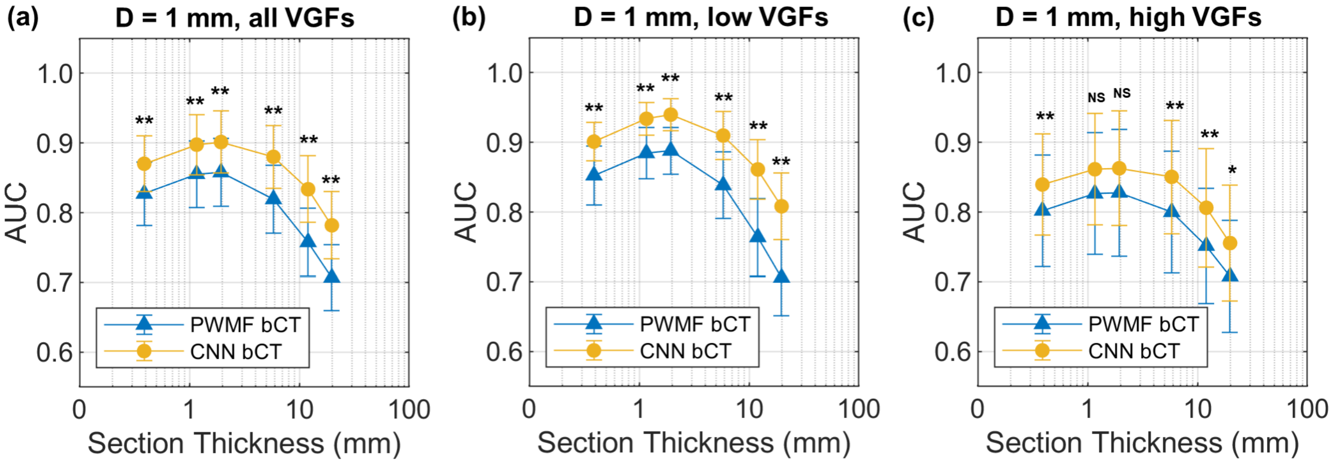

Figure 5.

Comparison of PWMF and CNN model observers on bCT images with 1 mm lesions as a function of section thickness across **(a)** all VGFs $(N = 24)$, **(b)** low VGFs $(N = 12)$, and **(c)** high VGFs $(N = 12)$. Paired t-tests were used with Bonferroni correction to adjust for multiple comparisons. One asterisk (*) is used to indicate $p < .05$, and two asterisks (**) are used to indicate $p < .01$. Error bars correspond to 95% confidence intervals for each performance estimate.

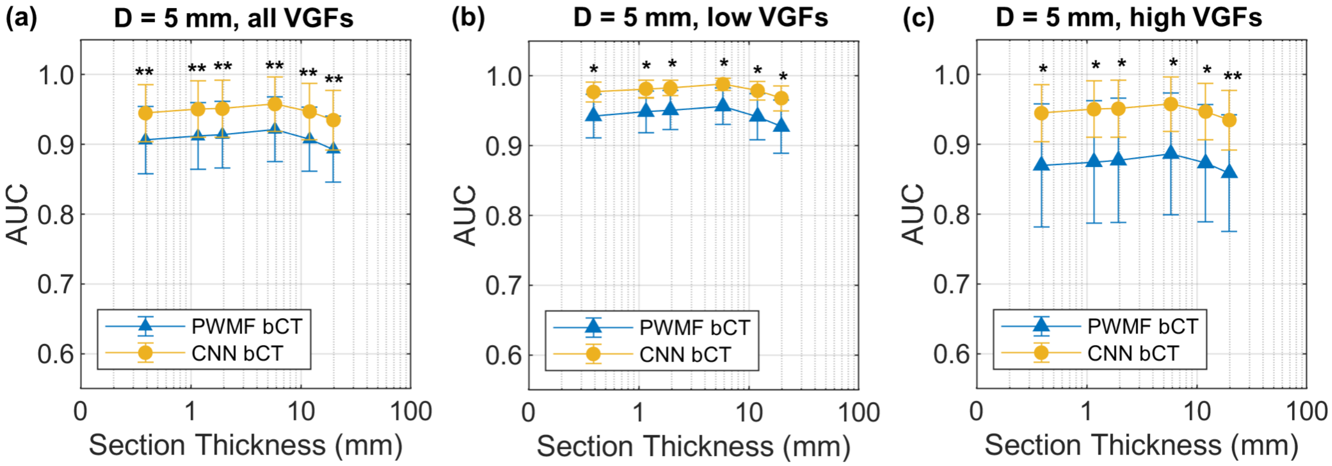

Figure 6.

Comparison of PWMF and CNN model observers on bCT images with 5 mm lesions as a function of section thickness across **(a)** all VGFs $(N = 24)$, **(b)** low VGFs $(N = 12)$, and **(c)** high VGFs $(N = 12)$. Paired t-tests were applied with Bonferroni correction to adjust for multiple comparisons. One asterisk (*) is used to indicate $p < .05$, and two asterisks (**) are used to indicate $p < .01$. Error bars correspond to 95% confidence intervals for each performance estimate.