JAMS
Journal of Associated Medical Sciences

Scopus Indexed & Thai-Journal Citation Index Centre (TCI)

Journal of Associated Medical Sciences

Journal homepage: https://www.tci-thaijo.org/index.php/bulletinAMS/index

Assessment of ACR phantom image quality in mammography using multiple Deep learning models

Thunyarat Chusin[1,2], Supitcha Saengkaew[1], Suradet Aunnoi[1], Amarin Thuikham[1], Pratchayakan Hompeng[1,3], Siriprapa Kaewjaeng[4,5], Titipong Kaewlek[1,2*]

[1]*Department of Radiological Technology, Faculty of Allied Health Sciences, Naresuan University, Phitsanulok Province, Thailand.*
[2]*Interdisciplinary Health and Data Sciences Research Unit, Faculty of Allied Health Sciences, Naresuan University, Phitsanulok Province, Thailand.*
[3]*Regional Medical Sciences Center 7 Khonkaen, Khonkaen Province, Thailand.*
[4]*Department of Radiologic Technology, Faculty of Associated Medical Sciences, Chiang Mai University, Chiang Mai Province, Thailand.*
[5]*Center of Radiation Research and Medical Imaging, Department of Radiologic Technology, Faculty of Associated Medical Sciences, Chiang Mai University, Chiang Mai Province, Thailand.*

ARTICLE INFO

Article history:
Received 16 August 2025
Accepted as revised 14 December 2025
Available online 17 December 2025

Keywords:
Mammography, ACR phantom, image quality assessment, deep learning, convolutional neural networks, artificial intelligence.

ABSTRACT

Background: Evaluating image quality in mammography—particularly using American College of Radiology (ACR) phantom images—is essential for maintaining diagnostic accuracy. Conventional evaluation relies on human visual inspection, which is prone to variability due to individual perception differences.

Objectives: This study examined the capability of multiple convolutional neural network (CNN)-based artificial intelligence (AI) models to assess the quality of ACR phantom images and address the limitations of human-based evaluation.

Materials and methods: Five CNN-based models—LeNet5, AlexNet, VGG19, GoogLeNet, and ResNet50—were used to classify 231 ACR phantom images acquired under different exposure settings. Dataset augmentation was performed by adding and removing artificial noise, increasing the dataset to 1,617 images. The dataset was then divided into training (70%), validation (10%), and testing (20%) subsets. Model performance was compared based on phantom image scoring.

Results: GoogLeNet showed the highest performance in evaluating fiber and mass groups, whereas ResNet50 achieved the best results for the speck group. The classification accuracy for scoring 16 object positions in ACR phantom images ranged from 31% to 100%, depending on object size.

Conclusion: To reliably classify acceptable mammographic image quality, model performance in detecting borderline cases (score 0.5) must improve. For clinical applicability, accuracy should exceed 80%.

Introduction

Mammography is a widely used imaging modality that employs low-dose X-rays for breast cancer screening and diagnosis. Women aged 40 and older are typically advised to undergo mammographic screening every 1-2 years.[1] Early detection through mammography reduces disease severity and mortality rates.[2] This highlights the importance of maintaining high image quality for effective breast cancer screening.

* **Corresponding contributor.**
Author's Address: *Department of Radiological Technology, Faculty of Allied Health Sciences, Naresuan University, Phitsanulok Province, Thailand.*
E-mail address: titipongk@nu.ac.th
doi: 10.12982/JAMS.2026.029
E-ISSN: 2539-6056

Consequently, dependable imaging systems and standardized image quality evaluation are essential components of quality control (QC) in mammography.[3]

Consistently high image quality is critical for accurate diagnosis. Routine quality control procedures are performed to ensure optimal system performance and to reduce diagnostic errors. A standard approach involves phantom testing using American College of Radiology (ACR) phantoms. These tests evaluate key image quality parameters, including density, contrast, uniformity, and artifact presence. However, these evaluations are generally performed through human visual inspection, which is time-consuming and inherently subjective. Variations in visual perception and professional experience can introduce inconsistencies in the results.

Artificial intelligence (AI) has increasingly transformed medical imaging. Deep learning methods, particularly convolutional neural networks (CNNs), have gained prominence with architectures such as LeNet, AlexNet, VGGNet, GoogLeNet, and Residual Network (ResNet). LeNet was originally designed for handwritten digit classification using the Modified National Institute of Standards and Technology database of 28×28 grayscale images.[4] Its relatively simple architecture is makes it suitable for low-complexity datasets and lightweight applications, including blood cell classification,[5] pneumonia detection,[6] and breast cancer histopathology analysis.[7] AlexNet demonstrated the power of deep CNNs in large-scale image classification.[8] Its innovations—such as deeper convolutional layers, ReLU activation, and dropout regularization—improved learning efficiency and reduced overfitting. In medical imaging, AlexNet has been applied to tasks such as chest disease classification (e.g., tumor vs. normal in chest images)[9] and brain tumor differentiation using magnetic resonance imaging (MRI).[10] VGGNet uses deep architectures composed of small 3×3 convolutional filters.[11] This structure allows effective hierarchical feature learning by stacking multiple layers. VGGNet has been widely adopted in transfer learning for medical imaging tasks, including breast cancer histopathology classification,[12] mammography interpretation,[13] and ultrasound image analysis,[14] and image quality assessment (IQA),[15] such as binary or multi-class scoring of ACR phantom images. GoogLeNet introduced Inception modules that integrate multiple convolutional filter sizes (1×1, 3×3, and 5×5) within a single module.[16] This design enables efficient multi-scale feature extraction with fewer parameters. As a result, GoogLeNet is effective in detecting microcalcifications and small lesions in mammograms[17] and is suitable for real-time classification due to its computational efficiency. ResNet introduced residual connections to address the vanishing gradient problem, allowing the training of very deep networks.[18] It has shown strong performance in complex medical image analysis,[11] including MRI-based studies,[19] cancer subtype classification from biopsy images,[20] and IQA tasks such as resolution assessment and phantom image scoring.[21]

Beyond abnormality detection in radiographic imaging,[22,23] CNNs have also demonstrated strong performance in radiological image quality assessment, including general radiography and mammography.[21,24-26] In one study, eight CNN architectures with 3–10 convolutional layers were trained to detect structures in ACR phantom images, achieving up to 95% classification accuracy using a six-layer model.[21] Another study modified the VGG16 architecture for ACR image scoring, reporting an F1-score of 0.69 for multi-class classification and an F1-score of 0.93 with an area under the receiver operating characteristic curve (AUC) of 0.97 for binary classification.[24]

This study aimed to evaluate and compare the performance of different CNN architectures in assessing ACR phantom image quality. Understanding their capabilities will support the development of AI-assisted tools that help reduce evaluator variability and improve consistency in mammography QC.

Materials and methods

Data collection

Phantom images were acquired using the Planmed Sophie Classic mammography system (Planmed Oy, Finland) equipped with a digital image receptor and the ACR accreditation phantom (Mammo 156 phantom, Gammex, USA). The phantom contains three test groups—fiber, specks, and masses—comprising 16 test objects, as illustrated in Figure 1. These objects differ in shape, diameter, and thickness and are designed to simulate common breast lesions observed in clinical mammograms. Specifically, the phantom includes 6 fibers (F1-F6) with diameters of 1.56, 1.12, 0.89, 0.75, 0.54, and 0.40 mm; 5 speck groups (S1-S5), each containing 6 specks with diameters of 0.54, 0.40, 0.32, 0.24, and 0.16 mm; and 5 mass objects (M1-M5) with decreasing diameters and thicknesses of 2.00, 1.00, 0.75, 0.50, and 0.25 mm.[27]

The phantom images were acquired using different tube voltage and tube current–time product settings controlled by the automatic exposure control system. Tube voltages ranged from 25 to 30 kVp (in 1 kVp increments), while tube current–time products ranged from 10 to 303 mAs, corresponding to seven optical density levels, as shown in Table 1. Each imaging condition was repeated three times. In total, 231 phantom images were collected, covering a wide range of image quality levels, from optimal (27 kVp, 71 mAs) to both lower and higher extremes.

Table 1. Number of ACR phantom images captured from the Planmed Sophie Classic mammography system with various exposure techniques.

Target/filter combination	Exposure techniques		Number of images
	kVp	mAs	
Mo/Mo	25	18-303	36
	26	10-225	39
	27	10-161	39
	28	10-114	39
	29	10-85	39
	30	10-67	39
Total			231

Note: Mo: molybdenum

Image augmentation

Artificial Gaussian noise was added to, and subsequently removed from, the original images using a median filter in ImageJ version 1.54g. This process was used to expand the dataset. Three empirically selected noise levels-10%, 30%, and 100%—were applied to simulate low, medium, and high noise conditions, respectively. In this context, the percentages represent the relative standard deviation parameter of the Gaussian noise within the software and do not be correspond to calibrated detector noise levels or specific clinical exposure settings.

For each of the 231 original images, three noise levels and two processing conditions (noisy and median-filtered) were generated, producing 1,386 augmented images (231×3×2). When combined with the original images (N=231), the total dataset consisted of 1,617 images. The dataset was then randomly divided into training (N=1,134; ~70%), validation (N=161; ~10%), and test (N=322; ~20%) subsets, as summarized in Table 2.

Table 2. Number of images used in the training, validation, and testing subsets.

	Number of images			
	ACR phantom	6 Fibers	5 Specks	5 Masses
Training	1,134 (~70%)	6,804	5,670	5,670
Validation	161 (~10%)	966	805	805
Testing	322 (~20%)	1,932	1,610	1,610
Total	1,617 (100%)	9,702	8,085	8,085
		25,872		

Image quality scoring by evaluators

This study was approved by the Institutional Review Board (IRB No. P1-0091/2567). All images in the training and validation subsets were scored by consensus between two researchers, while the testing subset was scored by consensus between two medical physicists. Each of the 16 phantom objects was assigned a score of 0, 0.5, or 1 according to the ACR digital mammography phantom scoring criteria.[28] All images were displayed on a 2-megapixel monitor, and the window width and level were adjusted for optimal viewing conditions based on evaluator judgment to ensure consistent and accurate scoring.

Image quality scoring by CNN-based AI models

All scored images in the training, validation, and testing subsets were cropped at the 16 predefined object locations, as shown in Figure 1. From each phantom image, 16 object-level images were extracted. This resulted in 18,144 training images (1,134×16), 2,576 validation images (161×16), and 5,152 testing images (322×16), as summarized in Table 2. Each object-level image was classified into one of three score categories (0, 0.5, 1) based on the corresponding image quality score. Five convolutional neural network (CNN)-based models—LeNet5,[4] AlexNet,[8] VGG19,[11] GoogLeNet,[16] and ResNet50[18]—were trained and validated for each of the 16 test objects. The LeNet5 model was trained from scratch using random weight initialization. In contrast, AlexNet, VGG19, GoogLeNet, and ResNet50 were initialized using pre-trained ImageNet weights and then fine-tuned on the phantom image dataset. This transfer learning strategy was employed to improve convergence of the deeper networks considering the relatively small dataset size. All CNN models were trained on Google Colab with a Tesla T4 GPU. A learning rate of 0.000001 and a batch size of 32 were applied consistently across all models. The number of training epochs was set to 20 for VGG19, 20 for GoogLeNet, 30 for AlexNet, 50 for ResNet50, and 60 for LeNet5. Training and validation loss and accuracy curves were generated for each object and model. These curves were used to assess network convergence and to confirm the optimization of training parameters, including batch size and number of epochs. During inference, the trained CNN models predicted image quality scores of 0, 0.5, or 1 for each test object.

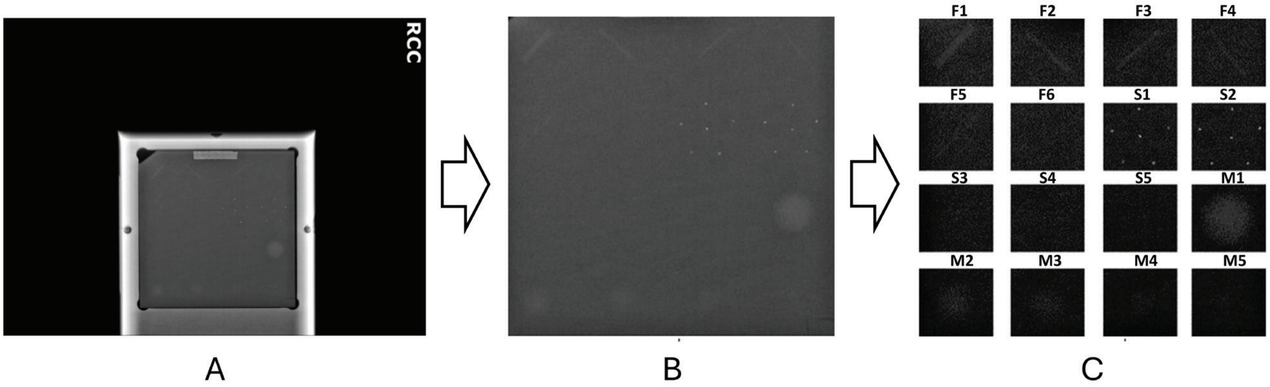


Figure 1. *A: original ACR phantom image, B: a cropped image of 1074x1044 pixels, C: A 268x261 pixel cropped image for the 16 test objects.*

Performance comparisons for CNN-based AI models

This study compared the performance of five CNN-based AI models: LeNet5, AlexNet, VGG19, GoogLeNet, and ResNet50. Model performance was evaluated using a multi-class confusion matrix, in which predicted scores were compared against three predefined image labels (0, 0.5, and 1) assigned by two medical physicists, as presented in Table 3.

Table 3. Analysis of the CNN-based AI models' performance based on multi-confusion matrix scoring.

Image scoring by CNN-based AI	Image scoring by medical physicists		
	1	0.5	0
1	TP_1, $TN_{0.5}$, TN_0	FP_1, $FN_{0.5}$, TN_0	FP_1, $TN_{0.5}$, FP_0
0.5	FN_1, $FP_{0.5}$, TN_0	TN_1, $TP_{0.5}$, TN_0	TN_1, $FP_{0.5}$, FP_0
0	FN_1, $TN_{0.5}$, FN_0	TN_1, $FN_{0.5}$, FN_0	TN_1, $TN_{0.5}$, TP_0

Where:

True Positive (TP): AI model correctly scored an image as 1 or 0.5, as marked by the medical physicists.

False Positive (FP): AI model incorrectly scored an image, assigning a higher score than the medical physicists.

False Negative (FN): AI model incorrectly scored an image, assigning a lower score than the medical physicists.

True Negative (TN): AI model correctly scored an image as 0, as marked by the medical physicists.

Performance of the five CNN-based AI models was then evaluated using accuracy, precision, sensitivity, specificity (recall), F1-score, and false positive rate (FPR), calculated using Equations (1) to (6) as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP+FP+FN+TN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

$$\text{F1 - score} = \frac{2 \text{ x precision x sensitivity}}{\text{precision + sensitivity}} \quad (5)$$

$$\text{False Positive Rate} = 1 - \text{Specificity} \quad (6)$$

Equation (7) was used to calculate the percentage accuracy of each model. Unlike confusion matrix-based metrics, accuracy is calculated by comparing the number of correctly predicted images to the total number of images, without distinguishing between the FP and FN.

$$\%\text{Accuracy} = \frac{\text{Number of corrected prediction image}}{\text{Total number of images}} \quad (7)$$

Results

The performance of five CNN-based AI models—LeNet5, AlexNet, VGG19, GoogLeNet, and ResNet50—was evaluated using key statistical metrics: accuracy, precision, sensitivity, specificity (recall), F1-score, and FPR. These results are presented in Figures 2-5. The analysis focused on classifying fibers, specks, and masses in ACR phantom images.

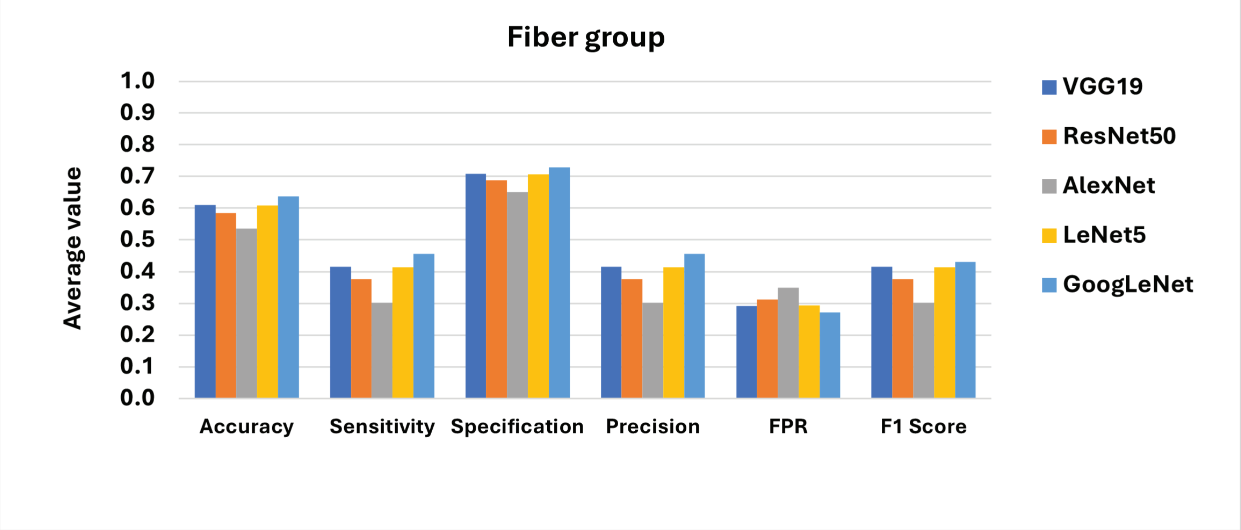


Figure 2. *Statistical metrics compared to all CNN-based AI models for classifying fiber objects.*

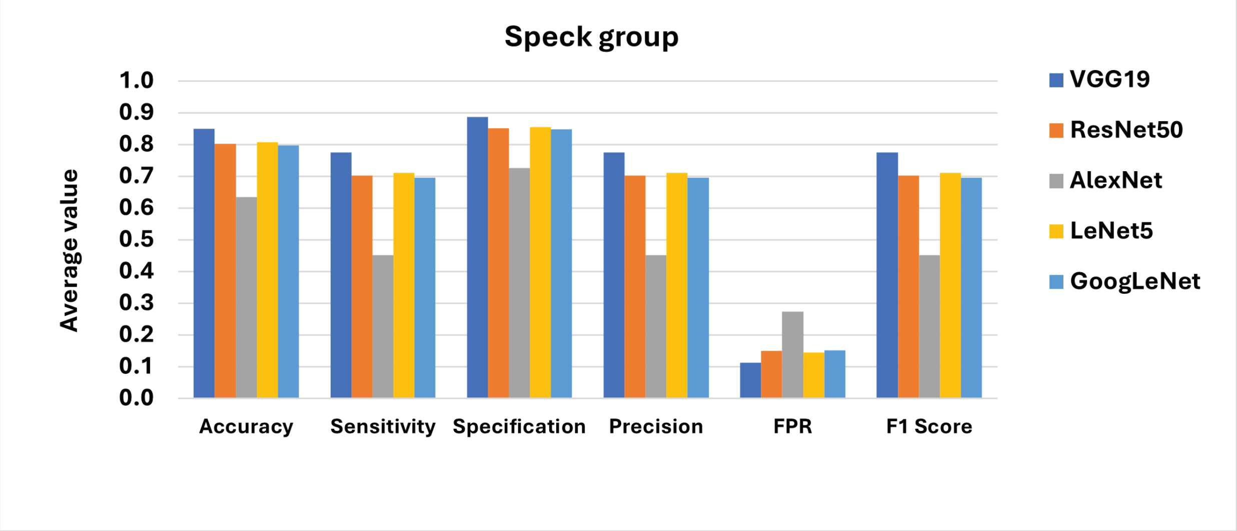


Figure 3. *Statistical metrics compared to all CNN-based AI models for classifying speck objects.*

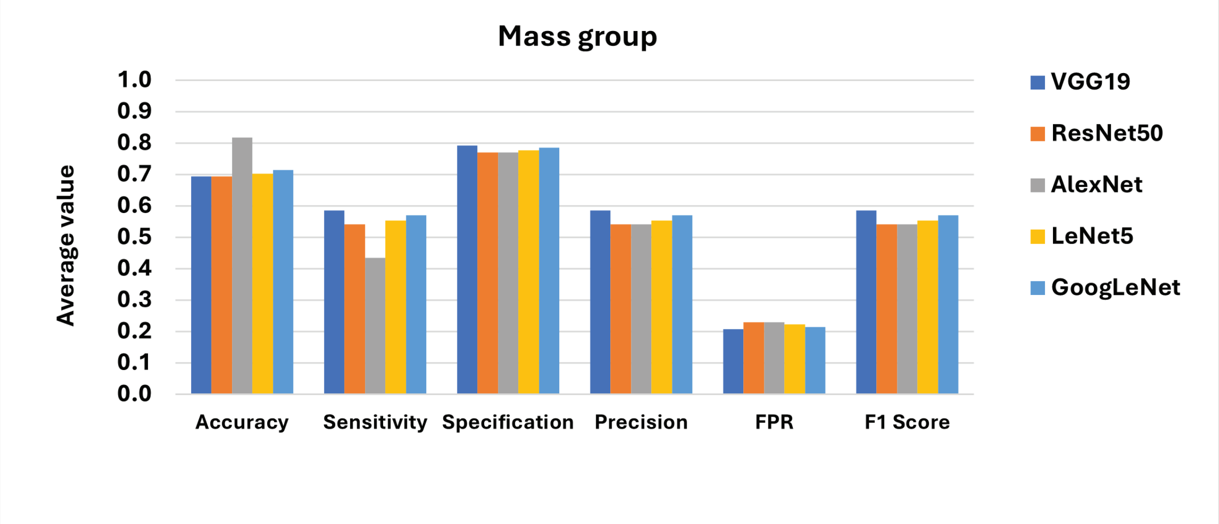
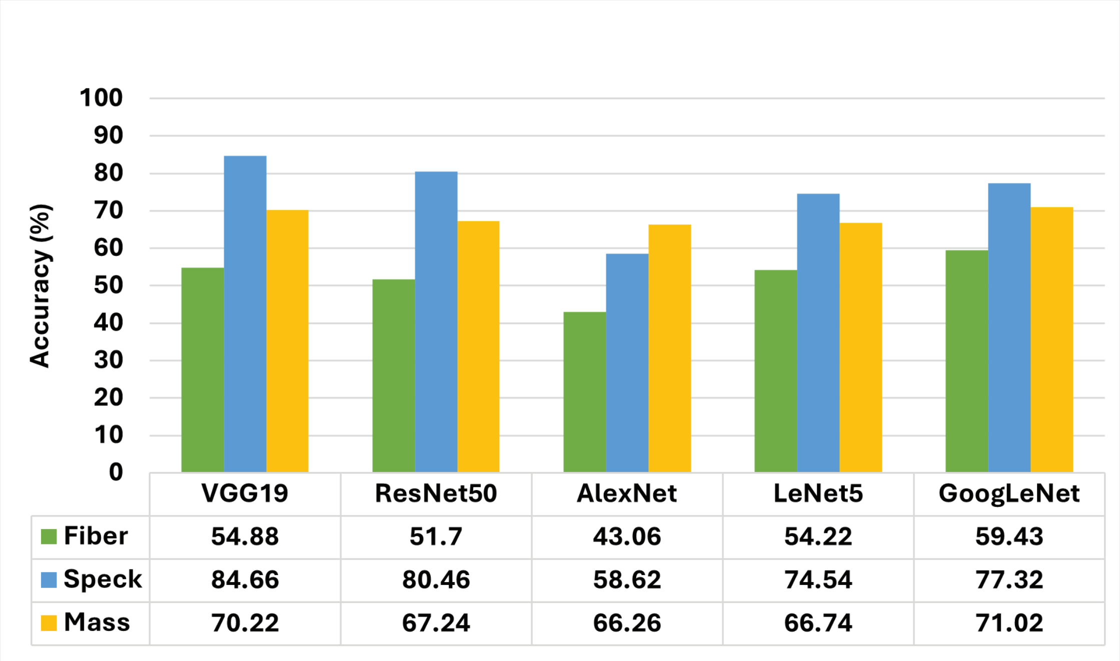


Figure 4. *Statistical metrics compared to all CNN-based AI models for classifying mass objects.*



	VGG19	ResNet50	AlexNet	LeNet5	GoogLeNet
🟩 Fiber	54.88	51.7	43.06	54.22	59.43
🟦 Speck	84.66	80.46	58.62	74.54	77.32
🟨 Mass	70.22	67.24	66.26	66.74	71.02

Figure 5. *Percentage accuracy compared to all CNN-based AI models for classifying fiber, speck and mass objects.*

The results showed that VGG19 and LeNet5 were unable to effectively differentiate object features, including size and density. Both models consistently produced the same output score of 1 for all test images, indicating a failure to distinguish among object classes. As a result, these two models were excluded as viable classifiers for fiber, speck, and mass objects in the phantom dataset.

The optimal CNN-based AI model for each object group was selected based on comparative statistical performance.

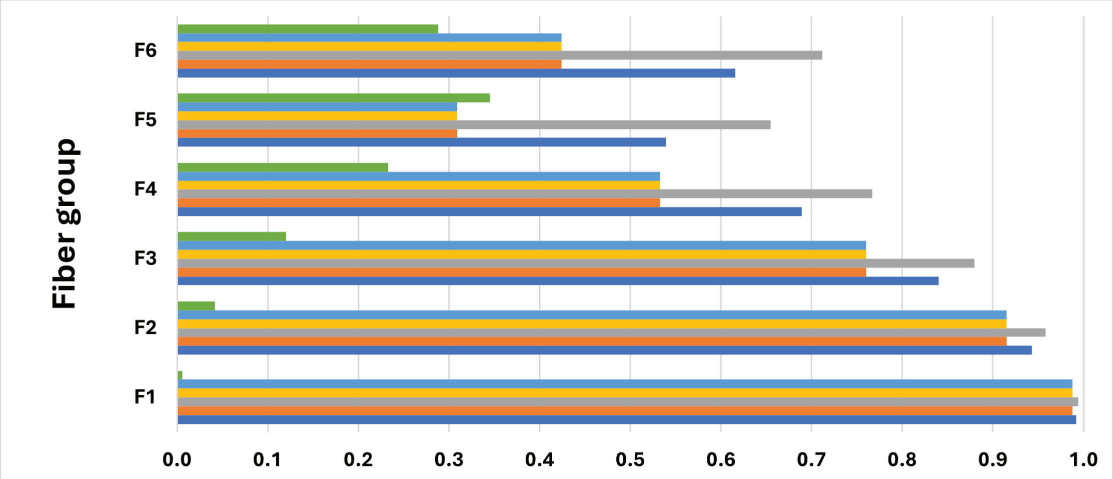
For the fiber group, GoogLeNet achieved the highest classification accuracy and underperformed the other models across most statistical metrics, while also demonstrating the lowest FPR. Therefore, GoogLeNet was identified as the most effective model for fiber evaluation.

For the speck group, although VGG19 and LeNet5 produced high statistical values, their ability to differentiate object features (due to identical output predictions) limited their usability. In contrast, ResNet50 demonstrated the highest accuracy, superior performance across key metrics, and the lowest FPR. Thus, ResNet50 was selected as the best model for speck classification.

For the mass group, VGG19 and GoogLeNet exhibited similarly high statistical values. However, VGG19 was excluded due to its inability to distinguish object features, consistently outputting a score of 1. Although AlexNet achieved the highest accuracy, its performance across other statistical metrics was inferior to GoogLeNet. Therefore, GoogLeNet was determined to be the most suitable model for mass object evaluation because it combined the highest accuracy with stronger overall metric performance and the lowest FPR.

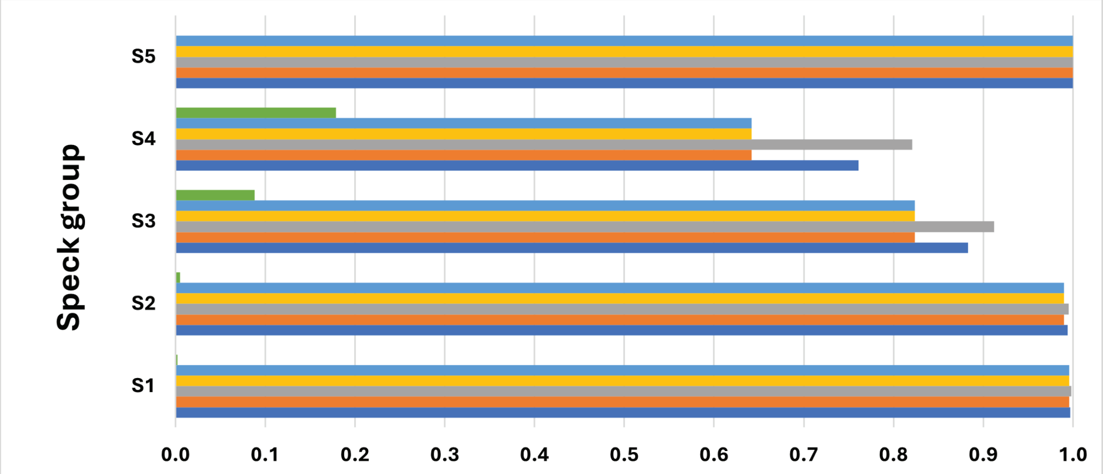
Figure 6 illustrates the performance of GoogLeNet in classifying fiber objects in ACR phantom images. In general, the FPR increased as fiber size and density decreased, except for F6, which had the smallest size and lowest density. In contrast, the other statistical metrics tended to decrease as fiber size and density decreased, except for F6. The GoogLeNet model demonstrated strong performance for F1, F2, and F3, with FPR values below 0.12 and other performance metrics exceeding 0.76.

Figure 7 shows the performance of the ResNet50 model in classifying speck objects. The model achieved FPR values below 0.09 and other statistical metrics greater than 0.82 for all speck groups, except S4.



	F1	F2	F3	F4	F5	F6
🟩 FPR	0.006	0.042	0.12	0.233	0.345	0.288
🟦 F1 Score	0.988	0.915	0.76	0.533	0.309	0.424
🟨 Precision	0.988	0.915	0.76	0.533	0.309	0.424
⬜ Specification	0.994	0.958	0.88	0.767	0.655	0.712
🟧 Sensitivity	0.988	0.915	0.76	0.533	0.309	0.424
🟦 Accuracy	0.992	0.943	0.84	0.689	0.539	0.616

Figure 6. *Performance of GoogLeNet model in classifying fiber objects in ACR phantom images.*

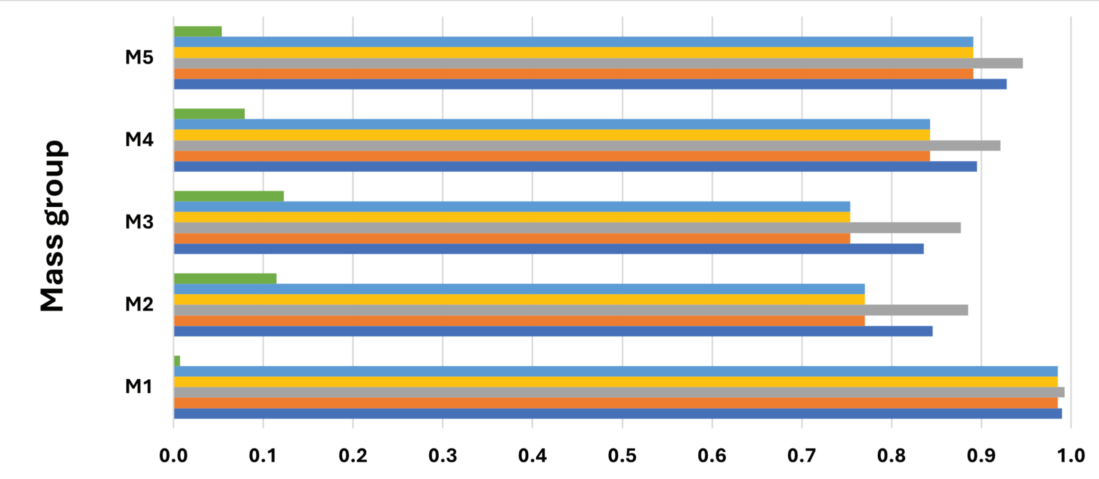


	S1	S2	S3	S4	S5
🟩 FPR	0.002	0.005	0.088	0.179	0.000
🟦 F1 Score	0.996	0.99	0.824	0.642	1.000
🟨 Precision	0.996	0.99	0.824	0.642	1.000
⬜ Specification	0.998	0.995	0.912	0.821	1.000
🟧 Sensitivity	0.996	0.99	0.824	0.642	1.000
🟦 Accuracy	0.997	0.994	0.883	0.761	1.000

Figure 7. *Performance of ResNet50 model in classifying speck objects in ACR phantom images.*

Figure 8 presents the performance of the GoogLeNet model in classifying mass objects. The FPR values were below 0.12, while the remaining statistical metrics exceeded 0.75 across all mass groups.

Figure 9 illustrates the classification accuracy of the GoogLeNet model for fiber and mass objects and the ResNet50 model for speck objects. The accuracy percentages for all 16 object positions were as follows: fibers-F1 (98.76%), F2 (91.51%), F3 (75.96%), F4 (53.33%), F5 (30.90%), and F6 (42.37%); specks-S1 (99.61%), S2 (99.03%), S3 (82.38%), S4 (64.20%), and S5 (100%); masses-M1 (98.53%), M2 (76.96%), M3 (75.37%), M4 (84.30%), and M5 (89.13%).



	M1	M2	M3	M4	M5
🟩 FPR	0.007	0.115	0.123	0.079	0.054
🟦 F1 Score	0.985	0.77	0.754	0.843	0.891
🟨 Precision	0.985	0.77	0.754	0.843	0.891
⬜ Specification	0.993	0.885	0.877	0.921	0.946
🟧 Sensitivity	0.985	0.77	0.754	0.843	0.891
🟦 Accuracy	0.99	0.846	0.836	0.895	0.928

Figure 8. *Performance of GoogLeNet model in classifying mass objects in ACR phantom images.*

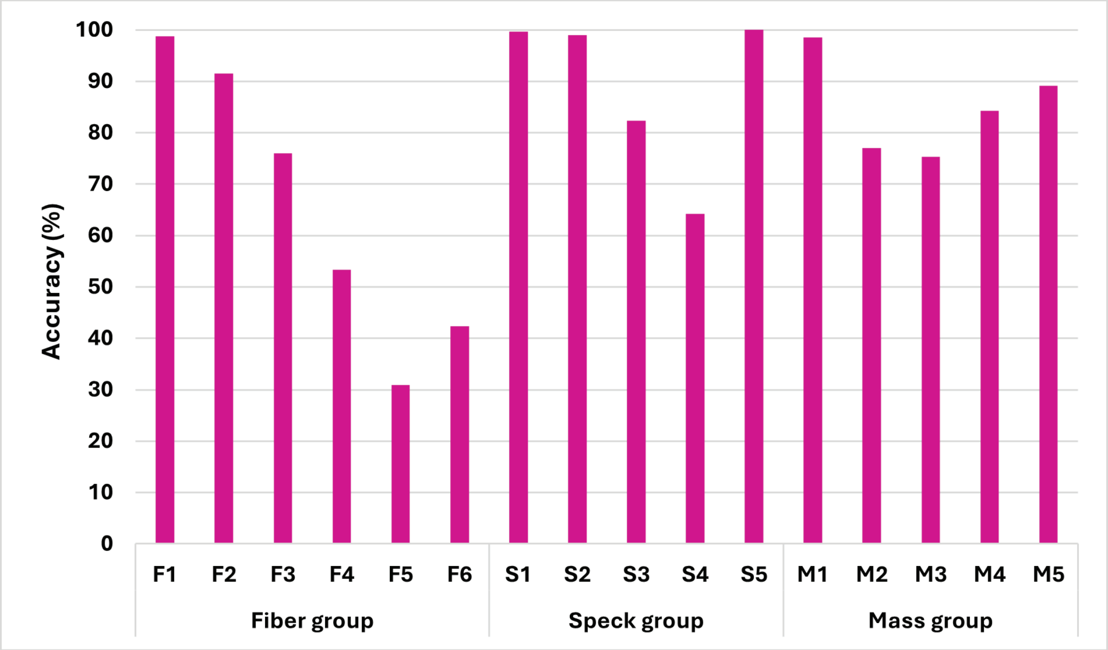


Figure 9. *Percentage accuracy of GoogLeNet in classifying fiber and mass objects, and the ResNet50 model in classifying speck objects.*

Discussion

This study evaluated the performance of five CNN-based AI models—LeNet5, AlexNet, VGG19, GoogLeNet, and ResNet50—in classifying fiber, speck, and mass objects in ACR phantom images. Model outputs were validated against object classifications performed by two medical physicists. The objects differed in shape, size, and density, which correspond to characteristics of typical lesions observed in clinical mammographic images.

VGG19 and LeNet5 failed to effectively differentiate object features because both models produced uniform outputs (score=1) across all test cases. This behavior indicates that the models did not learn meaningful decision boundaries. A possible cause is activation function saturation. Specifically, the original LeNet5 architecture uses the sigmoid activation function, which is prone to saturation. This leads to near-zero gradients, causing the vanishing gradient problem and limiting effective learning, particularly in deeper architectures. Furthermore, training VGG19 from scratch is known to be challenging. Issues such as improper weight initialization and unsuitable activation functions can prevent effective feature learning. These limitations may result in constant model outputs, indicating a failure to capture discriminative information from the input images.

According to ACR guidelines, mammographic image quality is considered acceptable when the visual image score reaches 4 or at least four fibers (F1-F4) are visible, a score of 3 or at least three specks (S1-S3) are identified, and a score of 3 or at least three masses (M1-M3) are detected, covering objects from large to small sizes. Based on the results of this study, the GoogLeNet model accurately classified F1, F2, and M1 with accuracy above 80%. The ResNet50 model effectively classified specks S1, S2, and S3, also exceeding 80% accuracy. These findings suggest that the prediction performance of GoogLeNet and ResNet50 must be improved, particularly for F3, F4, M2, and M3, to fully meet clinical image quality standards.

Although this study provides meaningful contributions, several limitations must be considered. First, all mammographic images were acquired using a single mammography system from one manufacturer with a Mo/Mo target/filter combination. Although data augmentation increased the dataset size, it did not add diversity in terms of vendors, system models, or beam qualities. Consequently, the reported performance of the CNN-based models may not generalize to phantom images acquired on other mammography systems or different target/filter combinations (e.g., Mo/Rh or Rh/Rh), which are commonly used in clinical quality control. To achieve robust and generalizable performance, the CNN models must be retrained or fine-tuned using phantom images acquired under clinically relevant exposure and target/filter configurations for each system.

A further limitation involves the imbalance in object score distribution. Images with an intermediate score of 0.5 were relatively scarce in both the training and test sets. This imbalance was most evident for smaller objects, which are less likely to be captured by the mammography system or reliably detected by human observers. Consequently, objects such as F4, F5, S4, and M4 often received intermediate scores of 0.5. In contrast, larger objects are more consistently captured and detected, leading to a score of 1 for F1, F2, M1, M2, S1, and S2. The smallest objects, although often not captured by the imaging system, were typically classified as non-visible and assigned a score of 0, as observed for F6, S5, and M5. Consequently, images with scores of 0 and 1 were more prevalent than those with a score of 0.5. In this study, scores of 1, 0.5, and 0 accounted for 56.5%, 19.0%, and 24.5% of all images, respectively, confirming that 0.5-score samples were significantly underrepresented.

All images, including those containing artificial Gaussian noise, were scored after data augmentation. The added noise can reduce the effective signal-to-noise ratio and, in some cases, shift clearly visible objects (score 1) into borderline cases (score 0.5), particularly for low-contrast fibers and masses. However, these borderline classifications were influenced not only by Gaussian noise but also by intrinsic object size and contrast, exposure conditions, and system noise, rather than by Gaussian noise alone. Another limitation is that data augmentation was limited to additive Gaussian noise followed by median filtering. Although this method directly targets quantum noise, it does not simulate other real-world variations, such as changes in positioning, geometric distortions, or subtle differences in object shape and contrast. Future studies should incorporate more advanced augmentation techniques, including small rotations and translations, elastic deformations, and controlled contrast or blur adjustments, to generate a more realistic and diverse training dataset.

As previously mentioned, CNN-based AI models have limited learning capacity when trained on under-represented image categories, particularly those assigned a score of 0.5. This class imbalance contributed to the reduced detection performance for F4, F5, S4, and M4. Increasing the number of images across all score categories, particularly for the 0.5 class, would likely improve model generalization and overall classification performance.

Sung Soo Park *et al.*[24] proposed a deep neural network-based phantom scoring method using the VGG16Net model for ACR phantom image quality evaluation. They applied two classification approaches: Multi-Class Classification (MCC), using scores of 0, 0.5, and 1, and Binary Classification (BCC), where scores of 0.5–1 were labeled as passing and 0 as failing. Their reported F1-score was 0.69 for MCC and 0.93 for BCC. This result indicates inferior performance for MCC, primarily due to class imbalance, especially the limited number of images with a score of 0.5. In the present study, the average F1-score for MCC was 0.592, which supports the same conclusion. The scarcity of 0.5-score images restricted model performance in both studies, emphasizing the need for a more balanced dataset.

Veli-Matti Sundell *et al.*[21] performed ACR phantom image quality scoring using a CNN-based AI model composed of six convolutional layers and achieved an overall accuracy of 95%. Their model was trained on a large dataset of 90,288 images, including 4,752 images with visible objects (score of 1) and 14,256 images with non-visible objects (score of 0), collected from eight mammography systems manufactured by three manufacturers. The substantial dataset size and system diversity enhanced model learning and improved generalization. However, their model sensitivity decreased as object size decreased, which is consistent with the reduced sensitivity for smaller objects observed in this study.

To address these limitations, future studies should use a larger and more diverse dataset that includes images of varying quality from multiple imaging systems to improve both model training and generalizability of CNN-based AI models. Beyond increasing dataset size, specific strategies are needed to enrich borderline cases (score of 0.5), which are critical for quality control decisions. Potential approaches include: (1) acquiring additional phantom images at exposure settings that produce near-threshold visibility for specific objects; (2) applying controlled contrast reduction and localized mild blurring to high-quality images to simulate subtle loss of conspicuity; (3) generating synthetic phantom images or using generative models to simulate near-threshold object visibility while preserving realistic background and noise characteristics; and (4) employing training strategies such as class-balanced sampling, focal loss, and cost-sensitive learning to reduce the impact of residual class imbalance during model optimization.

Conclusion

This study evaluated five CNN-based AI models for classifying objects in ACR phantom images. GoogLeNet demonstrated the best performance for fiber and mass classification, whereas ResNet50 was most effective for speck classification. GoogLeNet accurately classified fibers F1-F3 with an FPR below 0.12 and other statistical metrics above 0.76. It also demonstrated strong performance for mass classification, achieving an FPR below 0.12 and other statistical values exceeding 0.75. ResNet50 performed best for speck classification, with an FPR below 0.09 and other statistical metrics above 0.82, except for S4. Using the best-performing models across all 16 object positions, classification accuracies were as follows: fibers-F1 (98.76%), F2 (91.51%), F3 (75.96%), F4 (53.33%), F5 (30.90%), and F6 (42.37%); specks-S1 (99.61%), S2 (99.03%), S3 (82.38%), S4 (64.20%), and S5 (100%); and masses-M1 (98.53%), M2 (76.96%), M3 (75.37%), M4 (84.30%), and M5 (89.13%). Overall, GoogLeNet achieved accuracies ranging from 30.9% to 98.8% for fibers and masses, whereas ResNet50 achieved 64.2% to 100% for specks. To improve clinical applicability and align with ACR guidelines, future work should prioritize increasing prediction accuracy for intermediate objects, particularly for F3, F4, M2, and M3.

Ethical approval

This research was approved by the ethics committee of Naresuan University, Thailand (IRB No. P1-0091/2567.

Funding

This project is funded by the National Research Council of Thailand (NRCT) under contract number N42A670908 and awarded to Thunyarat Chusin.

Conflict of interest

This work has no conflicts of interest.

CRediT authorship contribution statement

Supitcha Saengkaew, Suradet Aunnoi, Amarin Thuikham, Pratchayakan Hompeng, and **Thunyarat Chusin**: methodology, investigation, data curation, and formal analysis; **Thunyarat Chusin**: original draft of the manuscript; **Siriprapa Kaewjaeng** and **Titipong Kaewlek**: supervision. All authors contributed to writing: review and editing, conceptualization and study design. All authors read and approved the final manuscript.

Acknowledgements

The authors acknowledge the Faculty of Allied Health Sciences, Naresuan University, for providing the mammography system used in this study.

References

[1] Calvocoressi L, Sun A, Kasl S V., Claus EB, Jones BA. Mammography screening of women in their 40s: Impact of changes in screening guidelines. Cancer 2008; 112: 473-80. doi: 10.1002/cncr.23210.

[2] Myers ER, Moorman P, Gierisch JM, Havrilesky LJ, Grimm LJ, Ghate S, et al. Benefits and harms of breast cancer screening: A systematic review. JAMA 2015; 314: 1615-34. doi: 10.1001/jama.2015.13183.

[3] Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition - Summary document. Ann. Oncol. 2008; 19: 614-22. doi: 10.1093/annonc/mdm 481.

[4] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. IEEE 1998; 86: 2278-324. doi: 10.1109/5.726791.

[5] Kousalya K, Krishnakumar B, Mohana RS, Karthikeyan N. Comparative analysis of White Blood Cells Classification using Deep Learning

Architectures. 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021; pp. 1220-5. doi: 10.1109/ICOSEC 51865. 2021.9591771.

[6] Jaganathan D, Balsubramaniam S, Sureshkumar V, Dhanasekaran S. Concatenated Modified LeNet Approach for Classifying Pneumonia Images. J Pers Med 2024;14: 328. doi: 10.3390/jpm14030328.

[7] Balasubramaniam S, Velmurugan Y, Jaganathan D, Dhanasekaran S. A modified LeNet CNN for breast cancer diagnosis in ultrasound images. Diagnostics 2023; 13: 2746 doi: 10.3390/diagnostics 13172746.

[8] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012; 25: 1097-105. doi: 10.1145/3065386

[9] Guo C, Chen Y, Li J. Radiographic imaging and diagnosis of spinal bone tumors: AlexNet and ResNet for the classification of tumor malignancy. J Bone Oncol 2024; 48: 100629. doi: 10.1016/j.jbo. 2024.100629.

[10] Azhagiri M, Rajesh P. EAN: enhanced AlexNet deep learning model to detect brain tumor using magnetic resonance images. Multimed Tools Appl 2024; 83: 66925-41. doi: 10.1007/s11042-024-181 43-w.

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2015 International Conference on Learning Representations (ICLR), 2015; pp. 1114. https://arxiv.org/abs/1409. 1556v6.

[12] Parvin F, Hasan MAM, Ahmed B, Mamun MA, Parvej SMK. Breast cancer histopathological image classification using an ensemble of deep convolutional neural networks. 2023 26th International Conference on Computer and Information Technology (ICCIT), 2023, pp. 1-5. doi. org/10.1109/ICCIT60459.2023.10441273.

[13] Harrison P, Hasan R, Park K. State-of-the-Art of Breast Cancer Diagnosis in Medical Images via Convolutional Neural Networks (CNNs). J Healthc Inform Res. 2023; 7: 387-432. doi: 10.1007/s416 66-023-00144-3.

[14] Chen Y, Shao X, Shi K, Rominger A, Caobelli F. AI in breast cancer imaging: An update and future trends. Semin Nucl Med. 2025; 55: 358-70. doi: 10.1053/j.semnuclmed. 2025. 01. 008.

[15] Amanova N, Martin J, Elster C. Explainability for deep learning in mammography image quality assessment. Mach Learn Sci Technol. 2022; 3: 025015. doi: 10.1088/2632-2153/ac7a03.

[16] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015; pp. 1-9. doi: 10.1109/ CVPR.2015.7298594.

[17] Teoh JR, Hasikin K, Lai KW, Wu X, Li C. Enhancing early breast cancer diagnosis through automated microcalcification detection using an optimized ensemble deep learning framework. PeerJ Comput Sci 2024;10:e2082. doi: 10.7717/peerj-cs. 2082.

[18] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, p. 770–8. doi: 10.1109/CVPR.2016.90.

[19] Zhang Y, Liu YL, Nie K, Zhou J, Chen Z, Chen JH, et al. Deep learning-based automatic diagnosis of breast cancer on MRI using mask R-CNN for detection followed by ResNet50 for classification. Acad Radiol. 2023; 30: S161-71. doi: 10.1016/j.acra. 2022.12.038.

[20] Tafavvoghi M, Sildnes A, Rakaee M, Shvetsov N, Bongo LA, Busund L-TR, et al. Deep learning-based classification of breast cancer molecular subtypes from H&E whole-slide images. J Pathol Inform 2025; 16: 100410. doi: 10.1016/j.jpi.2024. 100410.

[21] Sundell VM, Mäkelä T, Vitikainen AM, Kaasalainen T. Convolutional neural network -based phantom image scoring for mammography quality control. BMC Med Imaging 2022; 22: 216. doi: 10.1186/s12 880-022-00944-w.

[22] Al-antari MA, Al-masni MA, Choi MT, Han SM, Kim TS. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. Int J Med Inform. 2018; 117: 44-54. doi: 10.1016/ j.ijmedinf.2018.06.003.

[23] Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med. 2020; 121: 103792. doi: 10.1016/j.compbiomed.2020.103792.

[24] Park SS, Ku YM, Seo KJ, Whang IY, Hwang YS, Kim MJ, et al. Devising a deep neural network-based mammography phantom image filtering algorithm using images obtained under mAs and kVp control. Sci Rep. 2023;13: 3545. doi: 10.1038/s41598-023-30780-z.

[25] Ho PS, Hwang YS, Tsai HY. Machine learning framework for automatic image quality evaluation involving a mammographic American College of Radiology phantom. Phys Med. 2022; 102:1-8. doi: 10.1016/j.ejmp.2022.08.004.

[26] Tanaka R, Nozaki S, Goshima F, Shiraishi J. Deep learning versus the human visual system for detecting motion blur in radiography. J Med Imaging. 2022; 9: 015501. doi: 10.1117/1.jmi.9.1.015501.

[27] Song SE, Seo BK, Yie A, Ku BK, Kim HY, Cho KR, et al. Which phantom is better for assessing the image quality in full-field digital mammography?: American College of Radiology accreditation phantom versus digital mammography accreditation phantom. Korean J Radiol. 2012; 13: 776-83. doi: 10.3348/kjr.2012.13.6.776.

[28] Berns EA, Pfeiffer DE, Adent C, Baker JA, Bassett LW, Hendrick FRE, et al. Quality control manual for 2D and digital breast tomosynthesis. Radiologist's Section, Radiologic Technologist's Section, Medical Physicist's Section. American College of Radiology Subcommittee on Quality Assurance in Mammography of the Committee on Mammography Accreditation, American College of Radiology; 2018.