10.1148/radiol.2020200015

1. Appendix E1: Training Procedure

The deep learning (DL) model, described in the Materials and Methods Section, was a preactivation Resnet-34 network, where the batch normalization layers were replaced with group normalization layers (21–23). It was trained using the full-field digital mammography (FFDM) dataset (see Table 1) by use of the Adam optimizer (1) with a learning rate of 10^{-4} and a weight decay of 10^{-3}. Weight decay not was applied to the parameters belonging to the normalization layers. The input was resized to 416 × 320 pixels and the pixel intensity values were normalized so that the grayscale window denoted in the Digital Imaging and Communications in Medicine (DICOM) header ranged from 0.0 to 1.0. No additional preprocessing was performed. Training was performed using mixed precision (2) and gradient checkpointing (3) with batch sizes of 256 distributed across two NVIDIA GTX 1080 Ti graphics processing units (Santa Clara, CA). Each batch was sampled such that the probability of selecting a BI-RADS B or BI-RADS C sample was four times that of selecting a BI-RADS A or BI-RADS D sample, which roughly corresponds to the distribution of densities observed nationally in the United States (4). Horizontal and vertical flipping were employed for data augmentation. To obtain more frequent information on the training progress, epochs were capped at 100k samples compared with a total training set size of over 672k samples. The model was trained for 100 such epochs. Results are reported for the epoch that had the lowest cross entropy loss on the validation set, which occurred after 93 epochs.

Training from scratch on the synthetic 2D mammography (SM) datasets was performed following the same procedure as for the base model. For training from scratch, the size of an epoch was set to the number of training samples.

Appendix E2: Adaptation Methods

Three methods were employed to adapt the DL model from FFDM to SM and from Site 1 to Site 2: 1) vector calibration, 2) matrix calibration, and 3) fine-tuning.

The two calibration methods were originally proposed by Guo et al for the task for calibration and have been repurposed here for adaptation (25). In both methods, a linear operator is applied to the logits produced by the existing model. The parameters associated with the linear operator are learned as part of the adaptation process as described below. The updated probabilities predicted by the model are, then, given by:

$$p = \sigma\left(Az + b\right),$$

where z is the logits, A and b are the weights and bias associated with the linear operator, σ is the softmax operator, and p is the updated probabilities that an image belongs to each of the BI-RADS breast density categories. For vector calibration, A must be a diagonal matrix, while for matrix calibration, no restrictions are placed on A. The parameters for the vector and matrix calibration methods were chosen by minimizing a cross-entropy loss function by use of the BFGS optimization method (https://scipy.org, version 1.1.0). The parameters were initialized such that the linear layer corresponded to the identity transformation. Training was stopped when the ℓ_2-norm of the gradient was less than 10^{-6} or when the number of iterations exceeded 500.

For the fine-tuning method, no new layers or weights are introduced. Instead, a small portion of the network is retrained. In our case, only the last fully-connected layer is updated. For all layers, the model weights were initialized with the weights resulting from training on the FFDM dataset. Retraining the last fully-connected layer for the fine-tuning method was performed by use of the Adam optimizer with a learning rate of 10^{-4} and weight decay of 10^{-5}. The batch size was set to 64. The fully-connected layer was trained from random initialization for 100 epochs and results were reported for the epoch with the lowest validation cross entropy loss. For fine-tuning, the size of an epoch was set to the number of training samples.

Appendix E3: Density Distributions

The similarity between the DL model and radiologist density distributions is evaluated by use of several statistical techniques. Statistical significance for the difference between the radiologist and DL model density distributions is computed using a Pearson's χ^2 test. The similarity between the two distributions is also estimated using the Kullback-Liebler (KL) divergence where the radiologist distribution serves as the reference distribution. The 95% confidence intervals (CI) and variance of the KL divergence are estimated via bootstrapping (26). Significance for the relative similarity of two pairs of distributions, eg, the radiologist and DL model density distributions before and after adaptation, is estimated by comparing the KL divergences between the two pairs using a two-sided two-sample *t* test. Statements involving directional information, eg, "slightly underestimates," are evaluated based on a one-sided Wilcoxon signed-rank test.

The comparisons between the DL model and radiologist density distributions for the same dataset are summarized in Table E1. Comparisons of the relative similarity of the two distributions before and after adaptation (matrix calibration, 500 images) are also provided.

Appendix E4: Consistency of Image-level Predictions

The examination-level predictions of the DL model are obtained by averaging the predicted probabilities for the four BI-RADS breast density categories across all images in the examination. To better understand the consistency of the image-level predictions, the most probable BI-RADS breast density category for every image in an examination was considered for all examinations in the FFDM test set. Typically, the breast density predictions for different views within an examination were consistent. For example, the predictions were the same for all four views 79.5% (10546 of 13262) of the time. In almost all other cases, two BI-RADS breast density classes were predicted for the four views (20.5%, 2715 of 13272). In all but one case, the two predicted classes were neighboring density classes (eg, A and B). There was only one case where three distinct density predictions were made and no cases where four distinct predictions were made. In the case where three distinct density predictions were made, a prior surgery had removed most of the breast tissue from the right breast.

References

1. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y, eds. Presented at the third international conference on learning representations (ICLR). San Diego, Calif: May 7–9, 2015;

2. Micikevicius P, Narang S, Alben J, et al. Mixed precision training. Presented at the Sixth international conference on learning representations (ICLR), Vancouver, Canada, April 30–May 3, 2018.

3. Chen T, Xu B, Zhang C, Guestrin C. Training deep nets with sublinear memory cost. ArXiv 1604.06174 [preprint] https://arxiv.org/abs/1604.06174. Posted April 21, 2016.

4. Lehman CD, Arao RF, Sprague BL, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. Radiology 2017;283(1):49–58.

Table E1: A summary of the Breast Imaging Reporting and Data System (BI-RADS) breast density distributions of the DL model and the original reporting radiologists. Similarity between the distributions is characterized by the Kullback-Liebler (KL) divergence and Pearson's χ^2 test. Comparisons of the similarity before and after adaptation are calculated by comparing the KL divergences using a two-sided *t* test.

	Radiologist Distribution	DL Model Distribution	KL Divergence	*P* value
FFDM	A: 9.3%, B: 52.0%, C: 34.6%, D: 4.0%	A: 8.5%, B: 52.2%, C: 36.1%, D: 3.2%	0.0015 [0.0011, 0.0018]	< 0.001
Site 1 SM (Before)	A: 8.9%, B: 49.6%, C: 35.9%, D: 5.6%	A: 10.4%, B: 57.8%, C: 28.9%, D: 3.0%	0.02 [0.00, 0.03]	0.01
Site 1 SM (After)		A: 5.9%, B: 53.7%, C: 35.9%, D: 4.4%	0.008 [-0.005, 0.015]	0.24 (before vs after: 0.13)
Site 2 SM (Before)	A: 15.3%, B: 42.2%, C: 30.2%, D: 12.3%	A: 5.7%, B: 48.8%, C: 36.4, D: 9.4%	0.056 [0.041, 0.068]	< 0.001
Site 2 SM (After)		A: 16.9%, B: 43.3%, C: 29.4%, D: 10.4%	0.0026 [0.0011, 0.0035]	0.047 (before vs after: < 0.001)

Statistical significance was also calculated for other comparisons related to the density distributions. For the Site 1 SM test set, the DL model slightly underestimates the breast density relative to the radiologists ($P < .001$). For the Site 2 SM test set, the DL model did not underestimate the breast density ($P = .99$). The density distribution for the DL model for Site 2 is more similar to the radiologist distributions for Site 1 compared with the radiologist density distribution for Site 2 (Site 1 FFDM: KL = 0.03 [95% CI: 0.02, 0.05], $P = .03$; Site 1 SM: KL = 0.02 [95% CI:-0.02, 0.03], $P = .01$). This suggests that the model could have learned a prior density distribution from the Site 1 FFDM dataset that may not be optimal for Site 2 where patient demographics are different.