Automated Triage of Screening Breast MRI Examinations in High-Risk Women Using an Ensemble Deep Learning Model

Arka Bhowmik, PhD, Natasha Monga, MD,* Kristin Belen, MD,* Keitha Varela, BS,* Varadan Sevilimedu, DrPH,† Sunitha B. Thakur, PhD,*‡ Danny F. Martinez, MS,* Elizabeth J. Sutton, MD,* Katja Pinker, MD, PhD,* and Sarah Eskreis-Winkler, MD, PhD**

Objectives: The aim of the study is to develop and evaluate the performance of a deep learning (DL) model to triage breast magnetic resonance imaging (MRI) findings in high-risk patients without missing any cancers.

Materials and Methods: In this retrospective study, 16,535 consecutive contrast-enhanced MRIs performed in 8354 women from January 2013 to January 2019 were collected. From 3 New York imaging sites, 14,768 MRIs were used for the training and validation data set, and 80 randomly selected MRIs were used for a reader study test data set. From 3 New Jersey imaging sites, 1687 MRIs (1441 screening MRIs and 246 MRIs performed in recently diagnosed breast cancer patients) were used for an external validation data set. The DL model was trained to classify maximum intensity projection images as "extremely low suspicion" or "possibly suspicious." Deep learning model evaluation (workload reduction, sensitivity, specificity) was performed on the external validation data set, using a histopathology reference standard. A reader study was performed to compare DL model performance to fellowship-trained breast imaging radiologists.

Results: In the external validation data set, the DL model triaged 159/1441 of screening MRIs as "extremely low suspicion" without missing a single cancer, yielding a workload reduction of 11%, a specificity of 11.5%, and a sensitivity of 100%. The model correctly triaged 246/246 (100% sensitivity) of MRIs in recently diagnosed patients as "possibly suspicious." In the reader study, 2 readers classified MRIs with a specificity of 93.62% and 91.49%, respectively, and missed 0 and 1 cancer, respectively. On the other hand, the DL model classified MRIs with a specificity of 19.15% and missed 0 cancers, highlighting its potential use not as an independent reader but as a triage tool.

Conclusions: Our automated DL model triages a subset of screening breast MRIs as "extremely low suspicion" without misclassifying any cancer cases. This tool may be used to reduce workload in standalone mode, to shunt low suspicion cases to designated radiologists or to the end of the workday, or to serve as base model for other downstream AI tools.

Received for publication December 5, 2022; and accepted for publication, after revision, February 20, 2023.

From the Departments of *Radiology, †Epidemiology and Biostatistics, and ‡Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY.

Conflicts of interest and sources of funding: K.P. received payment for activities not related to the present article including lectures and service on speakers' bureaus and for travel/accommodations/meeting expenses unrelated to activities listed from the European Society of Breast Imaging (MRI educational course, annual scientific meeting), the IDKD 2019 (educational course), GBCC 10, TIBCS 2021, Olea Medical, Vara Merantix Healthcare GmbH, AURA Health Technologies GmbH, and Siemens Healthineers. She is also the Principal Investigator of a research project sponsored by Grail Inc, and a consultant for Merantix Healthcare, Siemens Healthineers, and Genentech, Inc. The other authors of this manuscript declare no conflicts of interest.

This work was supported in part by NIH/NCI Cancer Center Support Grant (P30 CA008748) and the NIH/NCI UG3 CA239861 grant.

Correspondence to: Sarah Eskreis-Winkler, MD, PhD, Department of Radiology, Memorial Sloan Kettering Cancer Center, 300 E 66th St, New York, NY 10065. E-mail: eskreiss@mskcc.org.

Supplemental digital contents are available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.investigativeradiology.com).

ISSN: 0020-9996/23/5810–0710
DOI: 10.1097/RLI.0000000000000976

Key Words: breast cancer, magnetic resonance imaging, high-risk, screening, deep learning, ensemble model

(*Invest Radiol* 2023;58: 710–719)

Each year, in the United States alone, there are over 300,000 cases of new breast cancer diagnoses, leading to over 40,000 deaths. Breast cancer mortality is significantly decreased by annual screening mammography, which is recommended in all average-risk women beginning at age 40 years.[1] For women at high risk of breast cancer, generally defined as a greater than 20% lifetime risk, screening guidelines recommend supplemental screening with dynamic contrast-enhanced breast magnetic resonance imaging (DCE-MRI), which markedly improves the sensitivity of early detection.[2] Dynamic contrast-enhanced MRI is also increasingly considered in women with a personal history of breast cancer and in women with dense breast tissue.[3–5] Recent, European Society of Breast Imaging guidelines recommend that women aged 50–70 years old with extremely dense breasts should be offered DCE-MRI every 2 to 4 years.[6]

As breast MRI utilization increases, it can be challenging for breast imaging radiologists to interpret all cases in a timely manner. Each screening breast MRI examination contains thousands of images for review, representing a nontrivial task. Yet, over 80% of these examinations are completely negative, requiring no further workup, and 98% to 99% are ultimately designated cancer-free after additional testing (eg, biopsy).[7] Given the limited number of radiologists with subspecialized training in breast imaging, there is a need for automated methods to optimize the clinical workflow. Automated triaging, for instance, could assign "extremely low suspicion" examinations to designated radiologists, or to be reviewed at the end of the workday, so that more time, focus, and expertise may be directed to the more challenging cases.

Over the past few years, deep learning (DL) tools have been developed for a variety of breast imaging applications,[8,9] including for image reconstruction,[10,11] segmentation,[12] cancer detection,[13] lesion classification,[14] and risk assessment.[15,16] Initial results are promising, with some algorithms performing at or beyond the level of radiologists,[17–19] although peer-reviewed studies have thus far been retrospective or small reader studies. Prospective trials are needed to more accurately determine the real-world performance and value of AI algorithms in the clinic.[20] Deep learning tools are particularly well suited for high-volume repetitive tasks such as cancer screening or examination triage, since they are not susceptible to fatigue that can lead radiologists to make errors in these setting. Several studies have used large mammography data sets to explore the use of DL to reduce the clinical workload via the automated interpretation of normal mammograms.[21–24] Deep learning has also been explored to reduce the clinical workload of screening breast MRI examinations in women with extremely dense breasts,[25] or as an initial triage step before computer-aided detection in an effort to decrease the number of biopsies on benign lesions.[26] Deep learning has also been used to triage cancer patients with additional MRI lesions directly to surgery, potentially eliminating need for additional preoperative biopsies that have a very high likelihood of malignancy.[27] However, no study to date has evaluated the use of DL for automated triage of screening breast

MRI examinations in a large population of high-risk women (ie, women with greater than 20% lifetime risk of breast cancer), which constitutes the clinically relevant scenario.

Herein, we aimed to develop and evaluate a DL model for the triage of "extremely low suspicion" MRI examinations in a large population of high-risk women. We further aimed to evaluate DL model performance in the clinical context, by comparing the performance of the DL model to that of fellowship-trained breast imaging radiologists.

MATERIALS AND METHODS

Study Patients

This retrospective study was approved by the institutional review board at Memorial Sloan Kettering Cancer Center, and the need for informed consent was waived. The study was compliant with the US Health Insurance Portability and Accountability Act.

Consecutive contrast-enhanced screening axial breast MRI examinations, performed between January 2013 and January 2019 across 6 Memorial Sloan Kettering Cancer Center cancer care sites in the states of New York and New Jersey were reviewed. Examinations were excluded; these are as follows: (1) if the postcontrast sequences were acquired in the sagittal plane or using a combined high spatial and high temporal resolution protocol, that is, differential subsampling with Cartesian ordering[28]; (2) if Digital Imaging and Communications in Medicine (DICOM) header labels were nonstandardized; (3) if the examinations were performed directly after neoadjuvant chemotherapy or breast surgery; or (4) if automated extraction of pathology and/or laterality information from the electronic medical record failed. Patient age, date of MRI examination, American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) assessment category,[29] background parenchymal enhancement (BPE) category, and pathology (including laterality information) with date of pathology, when applicable, were automatically extracted with the aid of in-house natural language processing software from the electronic medical record for all patients. Women with breast implants, prior cancer history, and prior surgical history were all included. Screening examinations were defined as MRIs performed in patients without a prior diagnosis of breast cancer, or in patients who had previously completed treatment for breast cancer. Patients with "recently diagnosed cancer" were defined as patients with a breast cancer diagnosis, but who did not yet undergo surgical treatment or chemotherapy. Each examination was given 2 labels: a BIRADS label (extracted from the radiology report), and a pathology label. The pathology label was cancer if the patient was diagnosed with breast malignancy within 1 year of the breast MRI examination date. The pathology label was negative (ie, no cancer), if the MRI assessment was negative (ie, BIRADS 1 or BIRADS 2), and there was at least 1 year of follow-up without malignant pathology or cancer diagnosis.

For this study, breast MRI data were divided based on the location of imaging (ie, New York or New Jersey). Breast MRI examinations performed at the New York sites (main site: Manhattan, regional site 1: Westchester, and regional site 2: Suffern) were used for the reader study and for the training/validation data set. The reader study test data set comprised 80 examinations randomly selected from the NY cohort. The remaining NY examinations were randomly split into 5 folds for training and validation. Breast MRI examinations performed at the New Jersey sites (regional site 3: Basking Ridge, regional site 4: Bergen, and regional site 5: Monmouth) were sequestered as an external validation data set; these examinations were further subdivided into "screening examinations" and "examinations performed in recently diagnosed cancer patients." Any patients with examinations in the reader study test set or the external validation set were excluded from the training/validation data set.

MRI Acquisition

All patients underwent a contrast-enhanced breast MRI on a 1.5 or 3.0 T system (Discovery 750; GE Medical Systems, Waukesha, WI) with a dedicated 8- or 16-channel breast coil. The gadolinium-based contrast agent was administered at a concentration of 0.1 mmol gadobutrol per kg body weight (Gadavist; Bayer Healthcare Pharmaceuticals, Inc, Whippany, NJ), at a rate of 2 mL/s. The acquisition parameters for conventional steady-state DCE-MRI, were TR/TE = 7.9/4.3, flip angle = 12 degrees, in-plane spatial resolution = 1.0 × 1.0 mm, thickness = 1.0 mm, temporal resolution = ~90 seconds, and axial orientation.

DL Model Development

The DL model was developed with the intent to triage a subset of breast MRI examinations as "extremely low suspicion" without missing any cancers. The model was tested using a histopathology reference standard.

To teach the model how to identify an "extremely low suspicion examination" that achieved this task, a 2-step labeling process was implemented. During training, BI-RADS labeling (see Fig. 1) consisted of 2 separately trained CNNs that performed (1) whole-breast segmentation and (2) examination triage into "extremely low suspicion" and "possibly suspicious" categories. Precontrast and first postcontrast fat-saturated T1-weighted, images were used to generate axial subtraction images, which were used to create standard maximum intensity projections (MIPs), as well as upper slab, middle slab, and lower slab MIPs, generated from the upper, middle, and lower axial images, respectively.[30]

Training of U-Net

Standard MIPs and their matching segmentations were used to train a 2D U-Net to perform whole-breast segmentation (see Supplemental Digital Content 1: Text S1 and Fig. S1 for the U-Net architecture, http://links.lww.com/RLI/A804).[12,31] The U-Net was initially trained using 500 randomly selected MIPs, with their associated manual segmentations serving as the reference standard. This initially trained U-Net was then used to generate reference standard masks for an additional 1500 MIPs from the training data set, which were used to retrain the U-Net using all 2000 MIPs. This final U-Net was used to generate segmentations for all remaining axial MIPs (see Supplemental Digital Content 1: Supplementary Table S1, http://links.lww.com/RLI/A804). The U-Net output, a segmented breast MIP, was automatically divided into right and left MIPs, which served as input to the DL classifier network, described below. This preprocessing segmentation step was performed to improve signal-to-noise ratio by focusing the attention of the DL model on the region of interest (ie, the breast) and avoiding distraction from background noise or high-intensity features located outside of the breast.

DL Classifier

A VGG-16 binary classifier was developed to triage segmented single breast MIPs as "extremely low suspicion" or "possibly suspicious" (see Supplemental Digital Content 1: Supplementary Text S1 and Supplementary Fig. S2 for the VGG-16 architecture, http://links.lww.com/RLI/A804).[32] Each single breast MIP was assigned both a BI-RADS assessment category label (automatically extracted from radiology reports) and a pathology label (automatically extracted from pathology reports).

During model development, BI-RADS labels were used to sort examinations into the "extremely low suspicion" and "possibly suspicious" categories. Training/validation data set examinations with BI-RADS 1 or BI-RADS 2 assessments were assigned an "extremely low suspicion" label, and examinations with BI-RADS 3, BI-RADS 4, BI-RADS 5, or BI-RADS 6 assessments were assigned a "possibly suspicious" label. During model evaluation, pathology labels were used as the reference standard to evaluate model performance. For examinations with an "extremely low suspicion" label, left and right breasts were treated as independent training examples, doubling the data set size. For examinations with a "possibly suspicious" label, only the breast containing the imaging finding was used (the contralateral breast was not used since it was unknown whether it had "possibly suspicious" imaging findings or not, and this would introduce noise into model training). The classifier was then

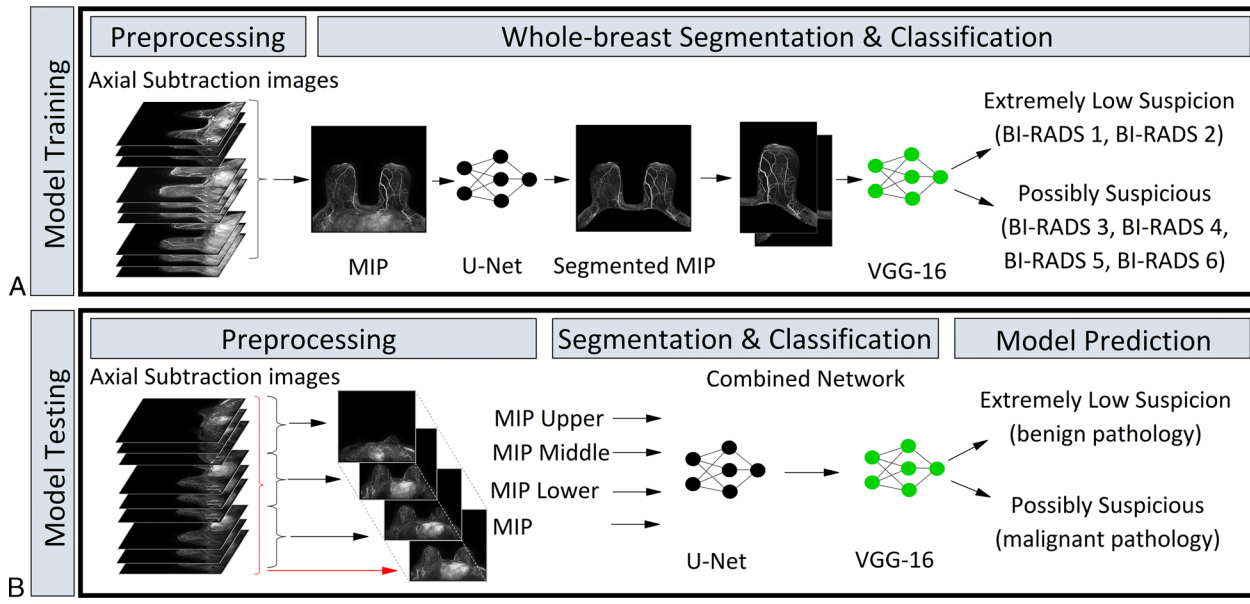

FIGURE 1. Flow diagram illustrating deep learning model triaging of breast magnetic resonance imaging examinations into "extremely low suspicion" and "possibly suspicious" examinations. A, Model training. B, Model testing. MIP, maximum intensity projection; BI-RADS, Breast Imaging Reporting and Data System.

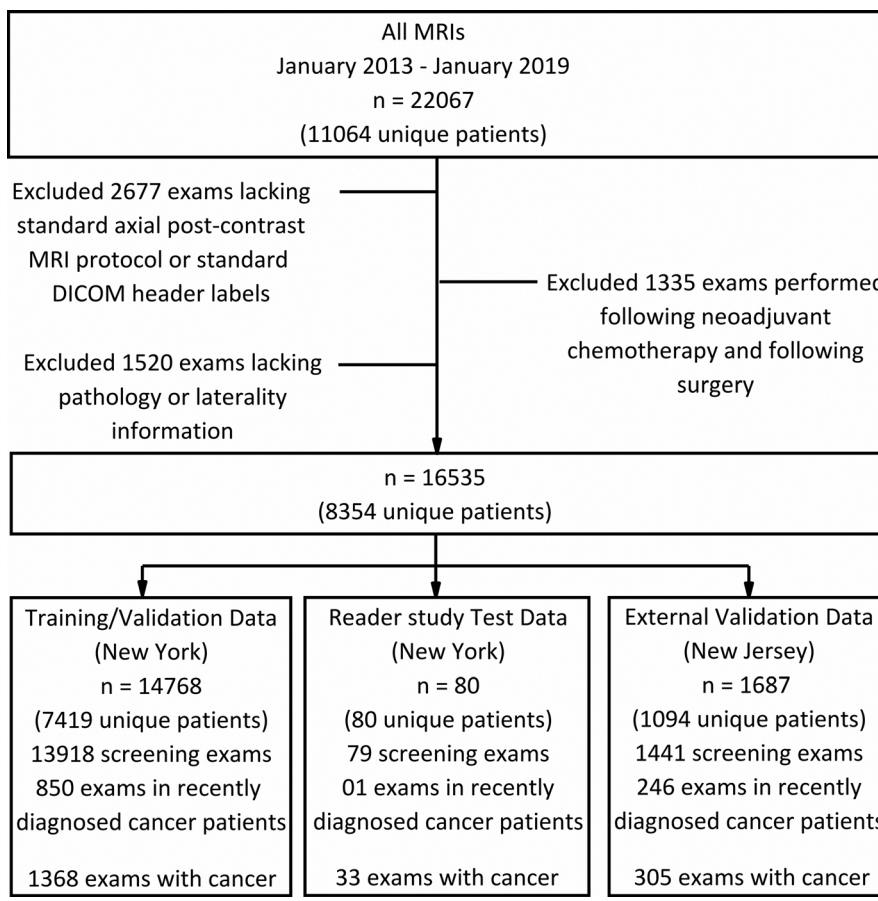

FIGURE 2. Flowchart illustrating the inclusion of breast magnetic resonance imaging examinations and patients.

initialized with ImageNet weights, and 5-fold cross-validation, was performed across all training/validation data, with class balancing using random oversampling of the minority class. The 5 classifiers were combined into an ensemble model, with the final prediction probability defined as the average of the 5 classifiers. To maximize sensitivity (ie, to avoid triaging any malignant cases as "extremely low suspicion"), the cutoff value of the final ensemble model was set as the probability score so that all malignant cases in all 5 validation folds to be classified as "possibly suspicious" (see Supplemental Digital Content 1: Supplementary Text S2 and Supplementary Fig. S3 for the calibration of the classifier, http://links.lww.com/RLI/A804).

For model testing, the final ensemble model was tested using the external validation data set consisting of examinations performed at the New Jersey sites, and using the reader study test data set. During model testing, after whole-breast segmentation, each subtraction MIP was used to generate 8 subimages: (1) left breast full MIP, (2) right breast full MIP, (3) left upper breast slab MIP, (4) left middle breast slab MIP, (5) left lower breast slab MIP, (6) right upper breast slab MIP, (7) right middle breast slab MIP, and (8) left lower breast slab MIP. Examinations were classified as "extremely low suspicion" if the predicted probability of all 8 subimages was less than the cutoff value that was established during model development.

All DL models were developed and tested in Python with Keras API and TensorFlow backend, using NVIDIA-GTX-1080ti GPU support. Hyperparameters for all models are detailed in the supplementary material (see Supplemental Digital Content 1: Supplementary Table S2, http://links.lww.com/RLI/A804). All code has been made available in GitHub (https://github.com/Arka-Bhowmik/mri_triage_normal).

Reader Study

To compare DL model performance to radiologist performance, 2 fellowship-trained breast imaging radiologists (N.M. and K.B., 5 and 10 years of experience, respectively) independently evaluated and assigned a suspicion score to the 80 breast MRI examinations in the reader study test

TABLE 1. Patient Demographical and Examination Characteristics

Parameters	Training and Validation Data Set	External Validation Data Set		Reader Study Test Data Set
		Screening Examinations	Examinations Performed in Recently Diagnosed Cancer Patients	
Examinations from New York imaging centers				
Main site: Manhattan	12,419	0	0	67
Regional site 1: Westchester	817	0	0	5
Regional site 2: Suffern	1532	0	0	8
Examinations from New Jersey imaging centers				
Regional site 3: Basking Ridge	0	976	157	0
Regional site 4: Bergen	0	62	25	0
Regional site 5: Monmouth	0	403	64	0
Totals from all imaging centers				
No. examinations	14,768	1441	246	80
No. patients	7419	874	246	80
Age distribution in years				
Mean	52	52	51	51
SD	11	10	10	11
Range	14–90	23–82	28–84	19–73
BI-RADS assessment				
BI-RADS 1	2934	202	0	11
BI-RADS 2	9276	910	0	36
BI-RADS 3	1074	146	0	0
BI-RADS 4	609	172	75	30
BI-RADS 5	25	11	0	2
BI-RADS 6	850	0	171	1
BPE				
Minimal BPE	5447	633	64	29
Mild BPE	4927	359	73	21
Moderate BPE	2444	156	59	21
Marked BPE	1007	47	8	8
BPE category (unknown)	943	246	42	1
Menopausal status				
Premenopausal	5870	606	117	33
Postmenopausal	8894	835	129	47
Menopausal status (unknown)	4	0	0	0
Pathology status				
Malignant	1368	59	246	33
Benign	13,400	1382	0	47

BI-RADS, Breast Imaging Reporting and Data System; BPE, background parenchymal enhancement.

data set. Readers were instructed that BI-RADS 1 and 2 assessments should be deemed "extremely low suspicion," and BI-RADS 3, BI-RADS 4, and BI-RADS 5 assessments should be deemed "possibly suspicious." Readers were blinded to the radiology reports, prior breast MRIs, and pathology information. Readers used all available MRI sequences to render their interpretation, mimicking routine clinical practice.

Gradient-Weighted Class Activation Mapping

Gradient-weighted class activation mapping (GRAD-CAM) is a technique to achieve explainable artificial intelligence (XAI) by visualizing a DL model's choice of features.[33] Herein, GRAD-CAM was used to highlight where the model was "looking." GRAD-CAM overlays were generated for each of the 5 DL classifiers.

Model Performance at Different Operating Thresholds

Because the clinical purpose of the DL model is to triage a subset a normal examination without missing cancer, the model threshold was set to maximize sensitivity, and the cutoff value was set as the probability score so that all malignant cases in all 5 validation folds during model development to be classified as "possibly suspicious." To explore the trade-off between "missed cancers" and accuracy, however, we evaluated model performance at a range of operating thresholds corresponding to the following sensitivities in the training/validation data set: 100%, 99%, 95%, 90%, 80%, and 0%.

Statistical Analysis

U-Net performance for whole-breast segmentation was evaluated using the Dice similarity coefficient (DSC), which measures the spatial overlap between the ground-truth segmented mask and the U-Net generated segmented mask.[34]

The DL model for triaging "extremely low suspicion" breast MRI examinations was evaluated using standard diagnostic performance metrics, including sensitivity and specificity, as well as a "workload reduction" metric, defined as the proportion of normal cases dismissed by the model as "extremely low suspicion." Calculations were performed with exact 95% confidence intervals (CIs) for the external validation data set and the reader study test data set.

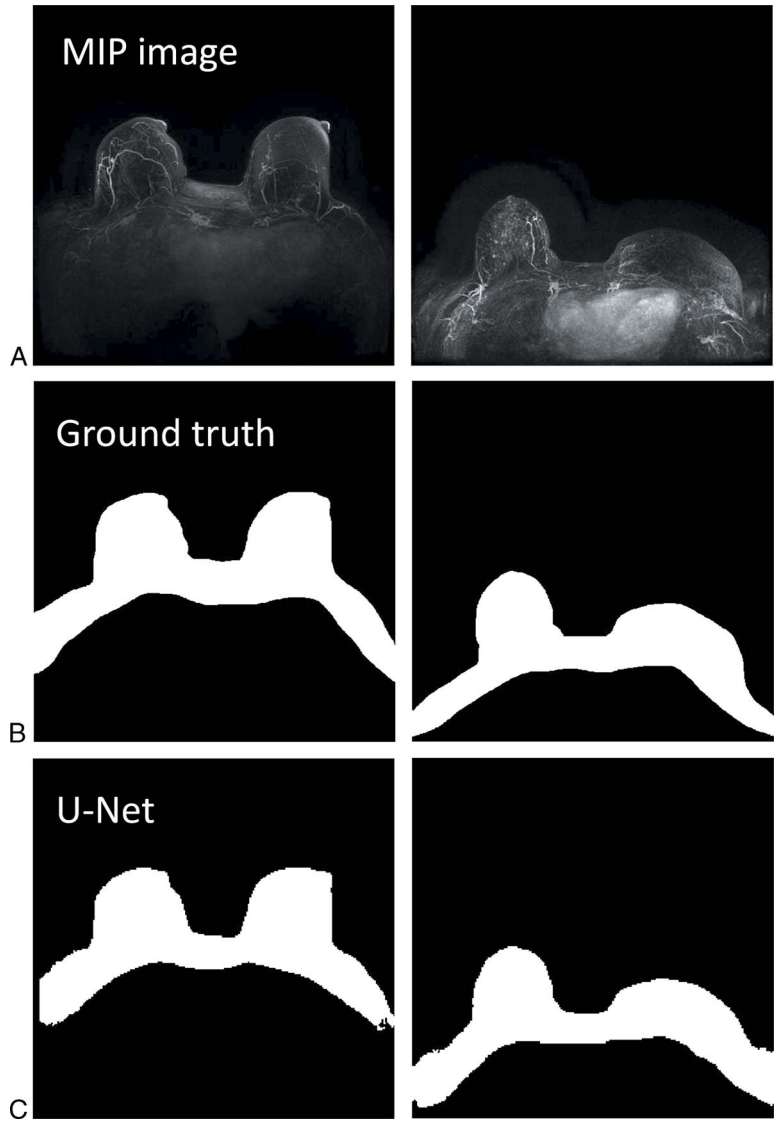

FIGURE 3. U-Net segmentation of the MIP image in 2 patients from the external validation data set. Input MIP (A), ground truth manual segmentation (B), and U-Net segmentation (C) (DSC: 0.928 and 0.934 for patient 1 and 2, respectively). MIP, maximum intensity projection.

Breast MRI examinations designated "extremely low suspicion" by the DL model were stratified according to the BI-RADS assessment category from the original radiology report, and also according to the BPE category from the original radiology report. Similar subgroup analysis was performed for MRI examinations designated "possibly suspicious" by the DL model.

RESULTS

Patient and Examination Characteristics

Of 22,067 breast MRI examinations performed in 11,064 patients between January 2013 and January 2019, 2677 were excluded because they were not acquired using a standard axial postcontrast MRI protocol or lacked standard DICOM header labels, 1335 were excluded because they were performed directly after neoadjuvant chemotherapy or after breast surgery, and 1520 were excluded due to the failure of automated extraction of pathology or laterality information. The study thus included a total of 16,535 breast MRI examinations performed in 8354 women (see Fig. 2 for the breast MRI and patient inclusion flow diagram).

Table 1 shows the patient demographical details and examination characteristics across the 3 data sets used in this study. The training and validation data set (from New York sites) consisted of 14,768 examinations performed in 7419 women (mean age, 52 years; range, 14–90 years). The external validation data set (from New Jersey sites) consisted of 1441 screening examinations performed in 874 women (mean age, 52 years; range, 23–82 years) and 246 examinations performed to evaluate the extent of disease in 246 women (mean age, 51 years; range, 28–84 years) with recently diagnosed cancer. The reader study test data set consisted of 80 women (mean age, 51 years; range, 19–73 years).

U-Net Performance for Whole-Breast Segmentation

U-Net whole-breast segmentation achieved a DSC of 0.94 (95% CI, 0.9368–0.9431) on 400 randomly selected examinations from the training and validation data set, a DSC of 0.93 (95% CI, 0.924–0.9329) on 120 randomly selected examinations from the external validation data set, and a DSC of 0.955 (95% CI, 0.9535–0.9583) on all 80 examinations in the reader study test data set (see Fig. 3 for 2 case examples).

DL Model Performance for Screening Breast MRI Examinations

For screening examinations in the external validation data set, the DL model triaged 159 (11.50%) of the 1382 cancer-free examinations to the "extremely low suspicion" category without missing a single cancer. It correctly classified all 59 (100%) of the malignant screening examinations as "possibly suspicious." This resulted in a sensitivity of 100%, a specificity of 11.50% (95% CI, 9.87–13.31), and a workload reduction of 11%.

Of the 159 screening breast MRI examinations that were correctly classified as "extremely low suspicion" by the DL model, 34 (21.38%) were BI-RADS 1 examinations, 114 (71.70%) were BI-RADS 2 examinations, 7 (4.40%) were BI-RADS 3 examinations, and 4 (2.52%) were BI-RADS 4 examinations. There were 115 (72.33%) examinations with minimal BPE, 26 (16.35%) with mild BPE, 5 (3.15%) with moderate BPE, 0 (0%) with marked BPE, and 13 (8.17%) with unknown BPE category.

Of the 59 screening examinations classified as "possibly suspicious" by the DL model, 168 (13.10%) were BI-RADS 1 examinations, 796 (62.10%) were BI-RADS 2 examinations, 139 (10.84%) were BI-RADS 3 examinations, 168 (13.10%) were BI-RADS 4 examinations, and 11 (0.86%) were BI-RADS 5 examinations. There were 518 (40.41%) of examinations with minimal BPE, 333 (25.97%) with mild BPE, 151 (11.78%) with moderate BPE, 47 (3.67%) with marked BPE, and 233 (18.17%) with unknown BPE category. See Figure 4 for further details. Table 2 illustrates how the DL model triaged the benign and malignant screening cases for each BI-RADS assessment category.

DL Model Performance for Breast MRI Examinations in Recently Diagnosed Cancer Patients

Of the 246 MRI examinations performed to evaluate extent of disease in recently diagnosed cancer patients, the DL model correctly classified all 246 (100%) malignant examinations as "possibly suspicious." See Table 3 for the breakdown by BI-RADS assessment category.

GRAD-CAM Results to Determine Where the DL Model Is "Looking"

GRAD-CAM maps were successfully reconstructed and showed that, in images containing malignancy, models from all 5 cross-validation

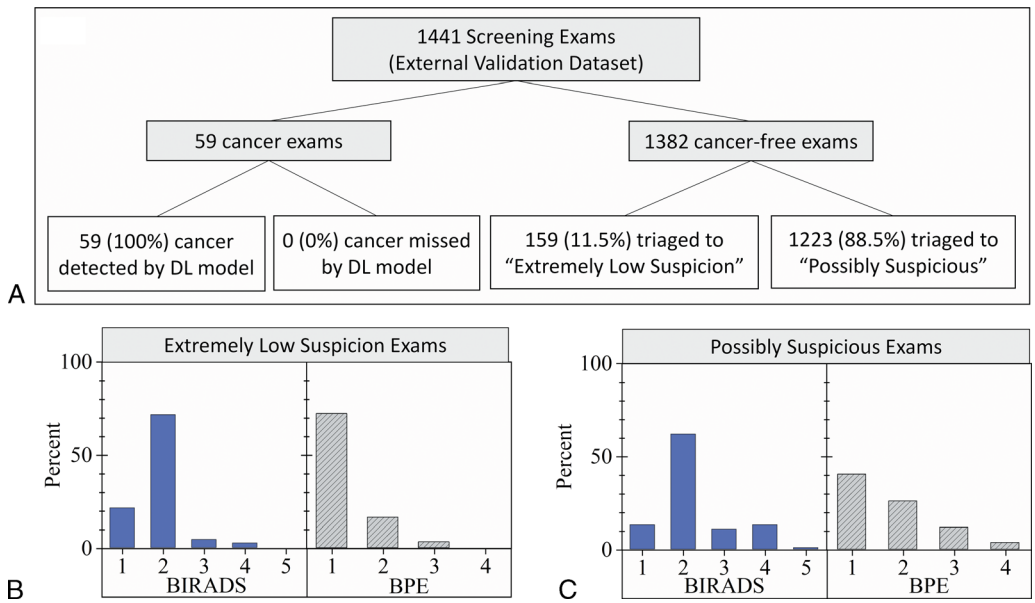

FIGURE 4. DL model triage of screening breast magnetic resonance imaging examinations from the external validation data set (A) with BI-RADS and BPE breakdown of examinations classified as "extremely low suspicion" (B) and "possibly suspicious" (C). DL, deep learning; BI-RADS, Breast Imaging Reporting and Data System; and BPE, background parenchymal enhancement.

TABLE 2. Screening Breast MRI Examinations (n = 1441) From the External Validation Data Set Triaged by the DL Model to "Extremely Low Suspicion" and "Possibly Suspicious" Examinations

			Screening Examinations, Triaged by the DL Model	
BI-RADS	**Pathology**	**Examinations**	**Extremely Low Suspicion**	**Possibly Suspicious**
1	Malignant	0	0 (0%)	0 (0%)
	Benign	202	34 (16.83%)	168 (83.17%)
2	Malignant	0	0 (0%)	0 (0%)
	Benign	910	114 (12.53%)	796 (87.47%)
3	Malignant	0	0 (0%)	0 (0%)
	Benign	146	7 (4.79%)	139 (95.21%)
4	Malignant	48	0 (0%)	48 (100%)
	Benign	124	4 (3.23%)	120 (96.77%)
5	Malignant	11	0 (0%)	11 (100%)
	Benign	0	0 (0%)	0 (0%)

MRI, magnetic resonance imaging; DL, deep learning; BI-RADS, Breast Imaging Reporting and Data System.

folds focused on the suspicious parts of the image. In contrast, for images without suspicious features, model localization varied across cross-validation folds (see Fig. 5).

Performance of the DL Model Compared With Breast Imaging Radiologists

Of the 33 malignant breast MRI examinations contained within the reader study test data set, the DL model flagged all 33/33 (100%) as "possibly suspicious," whereas the 2 readers flagged 33 (100%) and 32 (96.97%), respectively. One reader missed 1 cancer in an examination complicated by marked BPE (see Fig. 6 for the GRAD-CAM visualization of the cancer missed by reader 2).

Of the 47 benign examinations contained within the reader study test data set, the model classified 9 (19.15%) as "extremely low suspicion" and 38 (80.85%) as "possibly suspicious." Readers respectively classified 44/47 (93.62%) and 43/47 (91.49%) as "extremely low suspicion" and 3/47 (6.38%) and 4/47 (8.51%) as "possibly suspicious."

Thus, reader 1 attained a sensitivity of 100% and a specificity of 93.62% (95% CI, 86.63–100), and reader 2 attained a sensitivity of 96.97% (95% CI, 91.12–100) and a specificity of 91.49% (95% CI, 83.51–99.46). The DL model attained a sensitivity of 100% and a specificity of 19.15% (95% CI, 7.90–30.39) (see Table 4 for a summary of triaged examinations from the reader study test set by reader 1, reader 2, and the DL model).

Model Performance at Different Operating Thresholds

The performance metrics of the DL model vary depending on the operating threshold. At an operating threshold corresponding to 99% sensitivity in the training/validation set, workload reduction is 19.98%, sensitivity is 96.61% (95% CI, 91.99–100), and specificity is 20.83% (95% CI, 18.69–22.98) for the external validation set screening examinations. At an operating threshold corresponding to 95% sensitivity in the training/validation set, workload reduction is 44.55%, sensitivity is 83.05% (95% CI, 73.47–92.62), and specificity is 46.45% (95% CI, 43.82–49.08) for the external validation set screening examinations. See Figure 7 for graphical displays of this trade-off, for the external validation screening examination data set, and for the reader study test data set.

DISCUSSION

We developed an automated DL model to triage a subset of breast MRI examinations as "extremely low suspicion" in an effort to improve the clinical workflow without missing any cancers. In external validation, the DL model dismissed 159 (11%) of 1441 screening breast MRI examinations as "extremely low suspicion" without missing a single cancer; the model also correctly triaged all 246 breast MRIs in patients with recently diagnosed cancers as "possibly suspicious." In the reader study test data set, the DL model triaged 19.15% of examinations as "extremely low suspicion," again without missing a single cancer.

The primary goal of this DL model is to improve clinical workflow by triaging a subset of completely negative breast MRIs as "extremely low suspicion" without missing any cancers. To achieve this objective, we adopted a training strategy of labeling BI-RADS 1 and BI-RADS 2 examinations as "extremely low suspicion" (ie, no suspicious or indeterminate imaging findings) and BI-RADS 3, BI-RADS 4, BI-RADS 5 and BI-RADS 6 examinations as "possibly suspicious" (ie, examinations with indeterminate and/or malignant imaging findings). Our goal was to teach the model to be a high sensitivity/low risk model, pushing examinations with almost any kind of lesion to the "possibly suspicious"

TABLE 3. Breast MRI Examinations That Had Been Performed to Evaluate the Extent of Disease in Patients With Recently Diagnosed Breast cancer (n = 246) From the External Validation Data Set Triaged by the DL Model to "Extremely Low Suspicion" and "Possibly Suspicious" Examinations

			Examinations With Recently Diagnosed Cancer, Triaged by DL Model	
BI-RADS	**Pathology**	**Examinations**	**Extremely Low Suspicion**	**Possibly Suspicious**
4	Malignant	75	0 (0%)	75 (100%)
6	Malignant	171	0 (0%)	171 (100%)

MRI, magnetic resonance imaging; DL, deep learning; BI-RADS, Breast Imaging Reporting and Data System.

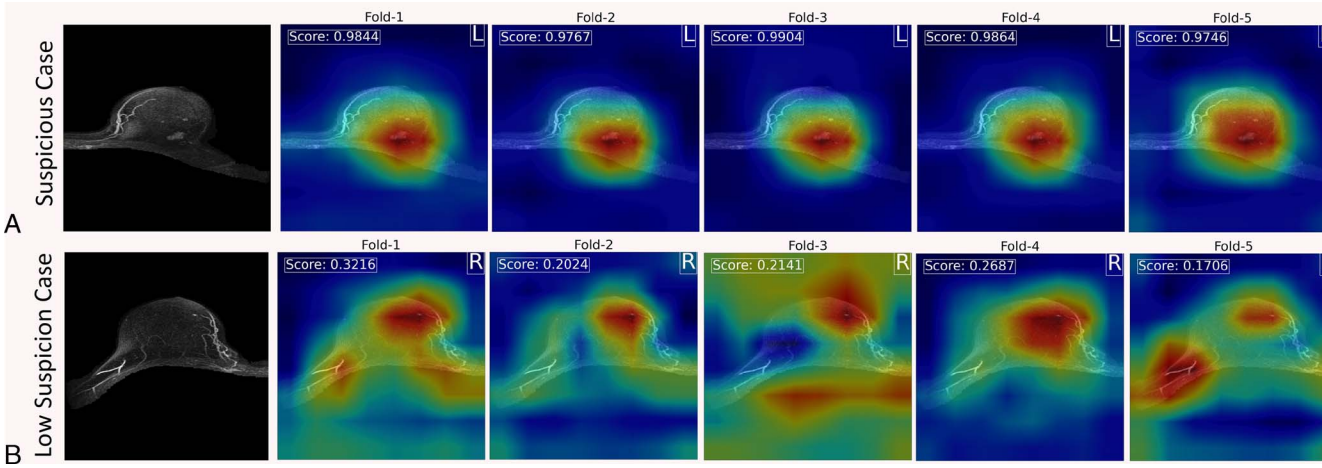

FIGURE 5. Visually explainable interpretation of DL model prediction with GRAD-CAM visualization for a case with left breast malignancy. A, In a cancer-containing examination, the ensemble model localizes to the suspicious mass in the left breast with all output scores above 0.97. B, For an "extremely low suspicion" examination without cancer, ensemble model localization varies across folds, with model output scores less than 0.33. GRAD-CAM, gradient-weighted class activation mapping; DL, deep learning.

category and sending only the most negative-appearing examinations to the "extremely low suspicion" category. However, BI-RADS labels are noisy and have high interreader variability. The algorithm, therefore, was ultimately tested not on BIRADS labels, but on a stronger pathology reference standard (based on a combination of pathology and 1 year of follow-up). Although a small number of cancers may be occult and not present for at least 1 year, the high sensitivity of MRI in the hands of fellowship-trained breast radiologists combined with the pathology information and 1 year follow-up yielded a strong reference standard for model evaluation in this work. This enabled calculation of the workload reduction to determine if any cancers were incorrectly dismissed by the model, which were the metrics that focused on the clinical motivation of the study, and to safely triage a fraction of "extremely low suspicious" examinations without missing any cancers. In addition, to further decrease the chances that cancers would be missed by the model, the operating threshold was adjusted to maximize the sensitivity of cancer detection across the 5 cross-validation folds (at the cost of reducing specificity).

Breast MRIs triaged as "extremely low suspicion" by our model were most likely to have been assigned BI-RADS 1 and BI-RADS 2 assessments and were also more likely to have been assigned low BPE (72.33% had minimal BPE and 16.35% had mild BPE). Of note, none of the examinations with marked BPE cases were triaged by the model as "extremely low suspicion." This suggests that, like radiologists, it is more difficult for the model to classify high BPE cases as normal. In the reader study, the malignancy missed by 1 of the 2 readers (but not by the model) had marked BPE.

This algorithm could be used to optimize the clinical workflow by triaging "extremely low suspicion" cases to designated radiologists, or to be reviewed the end of the workday when radiologist performance may be decreased. It is well documented that radiologist fatigue can affect diagnostic performance.[35] The algorithm could also be used to bypass, or at least shorten, radiologist interpretation of "extremely low suspicion" cases, although more extensive multicenter validation would be needed. In addition, it is important to note that if the algorithm was used as a standalone tool, an 11% reduction in number of cases to be read likely would not track linearly with time savings, since the simpler cases are dismissed by the model; future work is needed to better understand the time savings component.

This work adds to a growing body of evidence that DL models may be used to identify and triage completely normal-appearing examinations

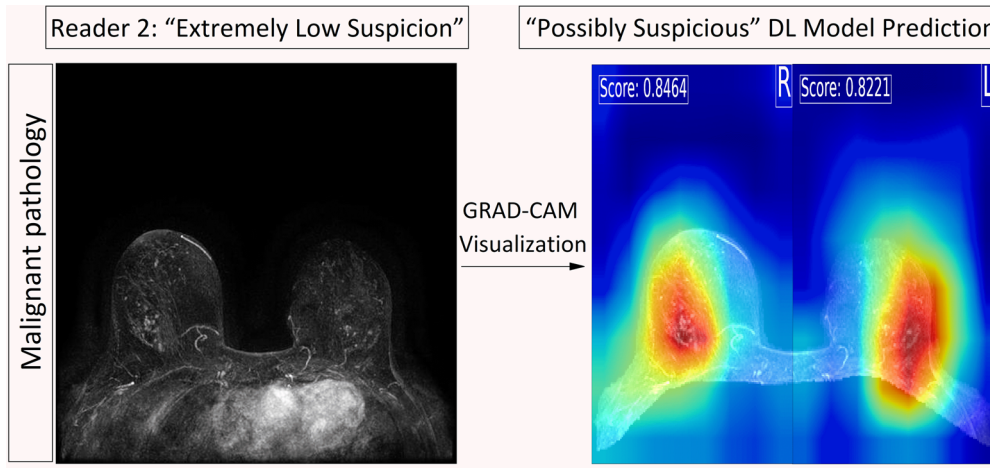

FIGURE 6. GRAD-CAM visualization of cancer examination missed by reader 2 in the reader study. GRAD-CAM, gradient-weighted class activation mapping; DL, deep learning.

TABLE 4. Triaged Examinations (n = 80) From the Reader Study Test Set

	Malignant		Benign	
Total No. Examinations	**33 (41.25%)**		**47 (58.75%)**	
Triaged Groups	**Extremely Low Suspicion**	**Possibly Suspicious**	**Extremely Low Suspicion**	**Possibly Suspicious**
Reader 1	0 (0%)	33 (100%)	44 (93.62%)	3 (6.38%)
Reader 2	1 (3.04%)	32 (96.96%)	43 (91.5%)	4 (8.5%)
DL model	0 (0%)	33 (100%)	9 (19.15%)	38 (80.85%)

DL, deep learning.

with high performance. Verburg et al[25] developed a DL algorithm to dismiss normal breast MRIs among non–high-risk women with dense breast tissue, achieving a workload reduction of 39.7%. The proportion of cancers in their cohort was much lower than in this study (1.6% cancers vs 18% cancers), likely accounting for their ability to dismiss a higher proportion of examinations. Jing et al[36] also developed a DL model to automatically dismiss 16% of normal ultrafast breast MRI examinations, albeit with only 30 cancers in their testing set. In triage studies for screening mammography, testing sets of >100,000 patients have been used to demonstrate a 25% decrease in callbacks with non-significant differences in overall diagnostic performance.[37]

Our study data were pooled from multiple imaging sites across 2 states, permitting us to sequester all data from one of the states as an external validation data set, thereby demonstrating the generalizability of our model. Our study was retrospective, but it is the largest study to date for the triage of normal breast MRIs in high-risk women, and it paves the way for future prospective work. All imaging sites used GE scanners, which may limit model generalizability. However, multiple versions of GE scanners were used, several MR protocol versions were used, and the technologists and patients varied by imaging site. Our DL model used 2D MIP images and MIP slabs down sampled to 256 × 256; in future work, we plan to extend the architecture to accommodate stacks of 3D subtraction images, dynamic contrast information, and other MR sequences as the model input, and explore the effect of increase input resolution to the original size of the MRI, thereby further improving the model's sensitivity for small or subtle findings. Our reference standard during training was the BIRADS label extracted from the radiology report. BIRADS labels have interreader variability, although we expect that

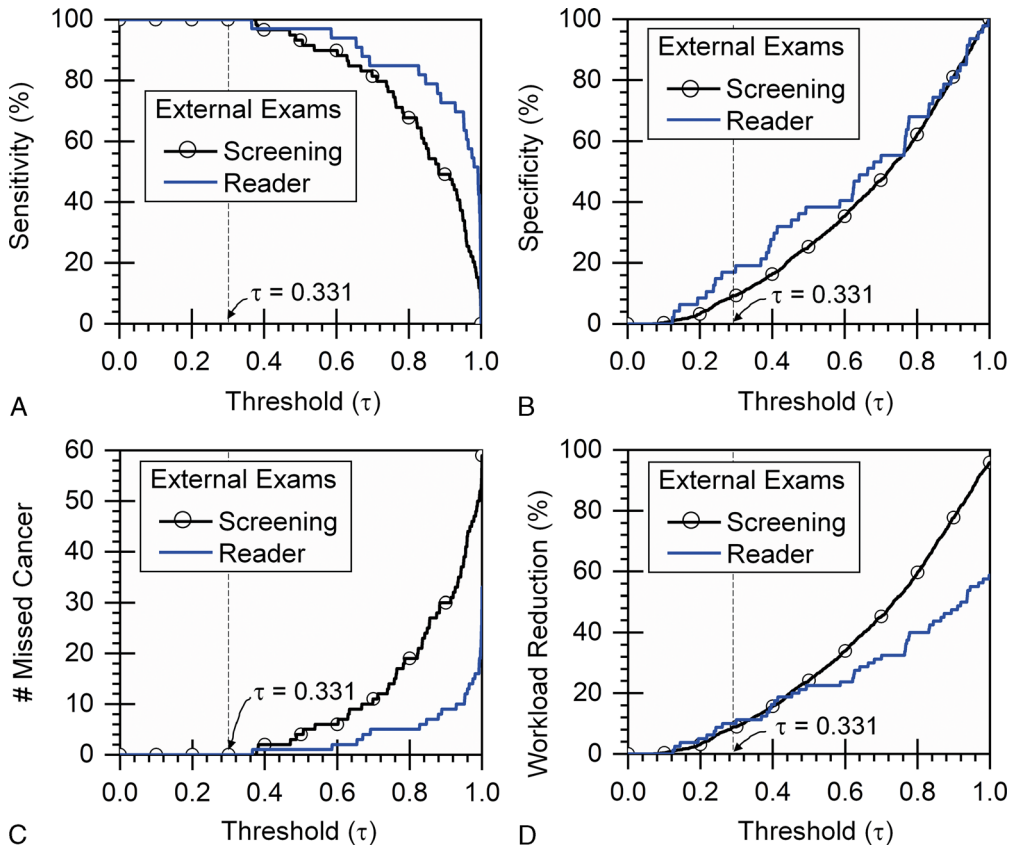

FIGURE 7. Ensemble DL model performance (A) sensitivity, (B) specificity, (C) missed cancer, and (D) workload reduction in external validation data set (ie, screening examinations and reader examinations) at different operating thresholds (τ). The operating thresholds were set using training/validation data. DL, deep learning.

the binary grouping of (1) BIRADS 1 and BIRADS 2, and (2) BIRADS 3, BIRADS 4, BIRADS 5, and BIRADS 6 mitigated some of this variability. The algorithm was then ultimately tested not on the BIRADS labels, but on a stronger reference standard of negative pathology combined with 1 year of negative follow-up. Although a small number of cancers may be occult and not present for at least 1 year, the high sensitivity of MRI in the hands of fellowship-trained breast radiologists combined with the pathology information and 1 year follow-up yielded a strong reference standard for model evaluation in this work.

In conclusion, our automated DL model safely triaged 11% of high-risk screening breast MRI examinations as "extremely low suspicion" without missing any of the 59 cancers among the 1441 screening breast MRI examinations in the external validation data set. This tool may be used to triage breast MRI examinations in clinic, shunting low suspicion cases to designated radiologists or to the end of a long workday. Multicenter prospective studies are warranted to pave the way for clinical implementation.

ACKNOWLEDGMENTS

The authors thank Joanne Chin, MFA, ELS, for her help with editing the manuscript, as part of her full-time employment at Memorial Sloan Kettering Cancer Center. The computation for this study was performed on the Lilac cluster hosted by the Sloan Kettering Institute in New York, NY.

REFERENCES

1. Expert Panel on Breast I, Mainiero MB, Moy L, et al. ACR Appropriateness Criteria® Breast Cancer Screening. *J Am Coll Radiol*. 2017;14(11S):S383–S390.
2. Monticciolo DL, Newell MS, Moy L, et al. Breast cancer screening in women at higher-than-average risk: recommendations from the ACR. *J Am Coll Radiol*. 2018;15(3 Pt A):408–414.
3. Bakker MF, de Lange SV, Pijnappel RM, et al. Supplemental MRI screening for women with extremely dense breast tissue. *N Engl J Med*. 2019;381:2091–2102.
4. Tilanus-Linthorst MM, Obdeijn IM, Bartels KC, et al. First experiences in screening women at high risk for breast cancer with MR imaging. *Breast Cancer Res Treat*. 2000;63:53–60.
5. Mann RM, Cho N, Moy L. Breast MRI: state of the art. *Radiology*. 2019;292:520–536.
6. Mann RM, Athanasiou A, Baltzer PAT, et al. Breast cancer screening in women with extremely dense breasts recommendations of the European Society of Breast Imaging (EUSOBI). *Eur Radiol*. 2022;32:4036–4045.
7. Lee JM, Ichikawa L, Valencia E, et al. Performance benchmarks for screening breast MR imaging in community practice. *Radiology*. 2017;285:44–52.
8. Hickman SE, Baxter GC, Gilbert FJ. Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. *Br J Cancer*. 2021;125:15–22.
9. Bhowmik A, Eskreis-Winkler S. Deep learning in breast imaging. *BJR|Open*. 2022;4:20210060.
10. Zhu B, Liu JZ, Cauley SF, et al. Image reconstruction by domain-transform manifold learning. *Nature*. 2018;555:487–492.
11. Kim M, Lee S-M, Park C, et al. Deep learning-enhanced parallel imaging and simultaneous multislice acceleration reconstruction in knee MRI. *Invest Radiol*. 2022;57:826–833.
12. Dalmis MU, Litjens G, Holland K, et al. Using deep learning to segment breast and fibroglandular tissue in MRI volumes. *Med Phys*. 2017;44:533–546.
13. Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans Med Imaging*. 2020;39:1184–1194.
14. Hu Q, Whitney HM, Li H, et al. Improved classification of benign and malignant breast lesions using deep feature maximum intensity projection MRI in breast cancer diagnosis using dynamic contrast-enhanced MRI. *Radiol Artif Intell*. 2021;3:e200159.
15. Yala A, Mikhael PG, Strand F, et al. *Sci Transl Med*. 2021;13:eaba4373.
16. Yala A, Mikhael PG, Strand F, et al. Multi-institutional validation of a mammography-based breast cancer risk model. *J Clin Oncol*. 2022;40:1732–1740.
17. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89–94.
18. Lotter W, Diab AR, Haslam B, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med*. 2021;27:244–249.
19. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst*. 2019;111:916–922.
20. Wallis MG. Artificial intelligence for the real world of breast screening. *Eur J Radiol*. 2021;144:109661.
21. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur Radiol*. 2019;29:4825–4832.
22. Shoshan Y, Bakalo R, Gilboa-Solomon F, et al. Artificial intelligence for reducing workload in breast cancer screening with digital breast tomosynthesis. *Radiology*. 2022;303:69–77.
23. Lang K, Dustler M, Dahlblom V, et al. Identifying normal mammograms in a large screening population using artificial intelligence. *Eur Radiol*. 2021;31:1687–1692.
24. Leibig C, Brehmer M, Bunk S, et al. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. *Lancet Digital Health*. 2022;4:E507–E519.
25. Verburg E, van Gils CH, van der Velden BH, et al. Deep learning for automated triaging of 4581 breast MRI examinations from the DENSE trial. *Radiology*. 2022;302:29–36.
26. Verburg E, van Gils CH, van der Velden BHM, et al. Validation of combined deep learning triaging and computer-aided diagnosis in 2901 breast MRI examinations from the second screening round of the dense tissue and early breast neoplasm screening trial. *Invest Radiol*. 2023;58:293–298.
27. Wang H, van der Velden BHM, Ragusi MAA, et al. Toward computer-assisted triaging of magnetic resonance imaging-guided biopsy in preoperative breast cancer patients. *Invest Radiol*. 2021;56:442–449.
28. Saranathan M, Rettmann DW, Hargreaves BA, et al. Variable spatiotemporal resolution three-dimensional Dixon sequence for rapid dynamic contrast-enhanced breast MRI. *J Magn Reson Imaging*. 2014;40:1392–1399.
29. D'Orsi CJ, Sickles EA, Mendelson EB, et al. *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*. Reston, VA: American College of Radiology; 2013.
30. Eskreis-Winkler S, Sutton EJ, D'Alessio D, et al. Breast MRI background parenchymal enhancement categorization using deep learning: outperforming the radiologist. *J Magn Reson Imaging*. 2022;56:1068–1076.
31. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer; 2015:234–241.
32. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*. 2014;1409.1556.
33. Groen AM, Kraan R, Amirkhan SF, et al. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: limited use of explainable AI? *Eur J Radiol*. 2022;157:110592.
34. Carass A, Roy S, Gherman A, et al. Evaluating white matter lesion segmentations with refined Sorensen-dice analysis. *Sci Rep*. 2020;10:8242.
35. Stec N, Arje D, Moody AR, et al. A systematic review of fatigue in radiology: is it a problem? *AJR Am J Roentgenol*. 2018;210:799–806.
36. Jing X, Wielema M, Cornelissen LJ, et al. Using deep learning to safely exclude lesions with only ultrafast breast MRI to shorten acquisition and reading time. *Eur Radiol*. 2022;32:8706–8715.
37. Lauritzen AD, Rodriguez-Ruiz A, von Euler-Chelpin MC, et al. An artificial intelligence-based mammography screening protocol for breast cancer: outcome and radiologist workload. *Radiology*. 2022;304:41–49.