Radiology

ORIGINAL RESEARCH • BREAST IMAGING

Lymph Node Metastasis Prediction from Primary Breast Cancer US Images Using Deep Learning

Li-Qiang Zhou, MD • Xing-Long Wu, PhD • Shu-Yan Huang, PhD • Ge-Ge Wu, MD • Hua-Rong Ye, MD • Qi Wei, MD • Ling-Yun Bao, MD • You-Bin Deng, MD, PhD • Xing-Rui Li, MD, PhD • Xin-Wu Cui, MD, PhD • Christoph F. Dietrich, MD, PhD

From the Sino-German Tongji-Caritas Research Center of Ultrasound in Medicine, Department of Medical Ultrasound, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, Hubei Province, China (L.Q.Z., G.G.W., Q.W., Y.B.D., X.W.C., C.F.D.); School of Mathematics and Computer Science, Wuhan Textile University, Wuhan, Hubei Province, China (X.L.W.); Department of Ultrasound, The First People's Hospital of Huaihua, University of South China, Huaihua, China (S.Y.H.); Department of Ultrasound, China Resources & Wisco General Hospital, Wuhan, Hubei Province, China (H.R.Y.); Department of Ultrasound, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China (L.Y.B.); Department of Thyroid and Breast Surgery, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei Province, China (X.R.L.); and Medical Clinic 2, Caritas-Krankenhaus Bad Mergentheim, Academic Teaching Hospital of the University of Wuerzburg, Bad Mergentheim, Germany (C.F.D.). Received February 20, 2019; revision requested April 2; revision received September 16; accepted September 26. **Address correspondence to** X.W.C. (e-mail: *cuixinwu@live.cn*).

Supported by the Young and Middle-aged Medical Talents Project of Wuhan, International Talent Cooperation Project of Henan (no. 2015GH7).

Conflicts of interest are listed at the end of this article.

See also the editorial by Bae in this issue.

Radiology 2020; 294:19–28 • https://doi.org/10.1148/radiol.2019190372 • Content codes: BR IN

Background: Deep learning (DL) algorithms are gaining extensive attention for their excellent performance in image recognition tasks. DL models can automatically make a quantitative assessment of complex medical image characteristics and achieve increased accuracy in diagnosis with higher efficiency.

Purpose: To determine the feasibility of using a DL approach to predict clinically negative axillary lymph node metastasis from US images in patients with primary breast cancer.

Materials and Methods: A data set of US images in patients with primary breast cancer with clinically negative axillary lymph nodes from Tongji Hospital (974 imaging studies from 2016 to 2018, 756 patients) and an independent test set from Hubei Cancer Hospital (81 imaging studies from 2018 to 2019, 78 patients) were collected. Axillary lymph node status was confirmed with pathologic examination. Three different convolutional neural networks (CNNs) of Inception V3, Inception-ResNet V2, and ResNet-101 architectures were trained on 90% of the Tongji Hospital data set and tested on the remaining 10%, as well as on the independent test set. The performance of the models was compared with that of five radiologists. The models' performance was analyzed in terms of accuracy, sensitivity, specificity, receiver operating characteristic curves, areas under the receiver operating characteristic curve (AUCs), and heat maps.

Results: The best-performing CNN model, Inception V3, achieved an AUC of 0.89 (95% confidence interval [CI]: 0.83, 0.95) in the prediction of the final clinical diagnosis of axillary lymph node metastasis in the independent test set. The model achieved 85% sensitivity (35 of 41 images; 95% CI: 70%, 94%) and 73% specificity (29 of 40 images; 95% CI: 56%, 85%), and the radiologists achieved 73% sensitivity (30 of 41 images; 95% CI: 57%, 85%; $P = .17$) and 63% specificity (25 of 40 images; 95% CI: 46%, 77%; $P = .34$).

Conclusion: Using US images from patients with primary breast cancer, deep learning models can effectively predict clinically negative axillary lymph node metastasis. Artificial intelligence may provide an early diagnostic strategy for lymph node metastasis in patients with breast cancer with clinically negative lymph nodes.

Online supplemental material is available for this article.

As the most common cancer among women worldwide, breast cancer poses a great challenge to public health on a global scale (1). Identification of the presence of lymph node metastasis is pivotal for the pathologic staging, prognosis, and guidance of treatment in patients with breast cancer (2). Although several histopathologic findings, such as vascular and lymphatic invasion, epithelial hyperplasia, and necrosis, are associated with a higher risk for lymph node metastasis, they are available only postoperatively (3). The preoperative prediction of lymph node metastasis can provide valuable information for determining adjuvant therapy and developing surgical plans, thereby facilitating pretreatment decisions.

Preoperative imaging assessment is of great value because of its convenient, comprehensive, and noninvasive properties. US plays a crucial role in detecting breast cancer and predicting lymph node metastasis (4). Most patients with early stage breast cancer who have clinically negative lymph nodes have no suspicious signs at either physical examination or imaging. Although radiologists often cannot find any signs of metastasis on US images of clinically negative lymph nodes, axillary lymph node metastasis is detected with sentinel lymph node biopsy in 15%–20% of patients (5). Several studies have found that numerous breast US characteristics are associated with lymph node metastasis. The distance

Abbreviations

AUC = area under the ROC curve, CI = confidence interval, CNN = convolutional neural network, ROC = receiver operating characteristic

Summary

The deep learning prediction model has the potential to predict lymph node metastasis in patients with clinically lymph node–negative breast cancer on the basis of US images of primary breast cancer.

Key Results

- For predicting lymph node metastasis on the basis of primary breast cancer US images, the deep learning model achieved an area under the receiver operating characteristic curve of 0.90 with the internal test set and 0.89 with the external test set.
- The three deep learning models—Inception V3, Inception-ResNet V2, and ResNet-101—achieved 85%, 78%, and 73% sensitivity (P = .40) and 73%, 75%, and 73% specificity (P = .96), respectively, in predicting lymph node metastasis with an independent external test set compared with 73% sensitivity (P = .51) and 63% specificity (P = .62) from a consensus of five radiologists.

of breast cancer from the skin and the nipple on US images is reported as a risk factor for lymph node metastasis (6). The presence of lymphatic invasion and the size of the primary tumor are also associated with lymph node metastasis (7). In addition, increased stiffness of the primary tumor as measured with shear wave elastography was associated with lymph node metastasis in patients with breast cancer (8).

Artificial intelligence, particularly deep learning algorithms, is gaining extensive attention for its excellent performance in image recognition tasks (9). Artificial intelligence models can find in medical images details that human experts cannot see and can automatically make a quantitative assessment. Deep learning algorithms have been widely applied in the field of image diagnosis and prediction owing to their advantages of being fast, accurate, and reproducible (10). The prediction of lymph node metastasis by combining the US signs of primary tumors and artificial intelligence may yield a great diagnostic effect. The aim of this study was to investigate the potential of deep learning algorithms for the prediction of clinically negative axillary lymph node metastasis through use of US images of primary breast cancer.

Materials and Methods

Patients and Data Sets

This retrospective multicohort study was approved by the institutional review board of Tongji Medical College of Huazhong University of Science and Technology, Hubei, China, and the requirement to obtain informed consent was waived (approval number: 2019S875). For the primary cohort, we assessed the Tongji Hospital medical records database from May 2016 to October 2018 to identify patients with histologically confirmed breast cancer who underwent surgical resection. Appendix E1 (online) presents the inclusion and exclusion criteria. In total, 680 patients comprised the training and validation sets (mean age, 48 years; range, 24–81 years) and 76 comprised the internal test set (mean age, 50 years; range, 25–82 years). From October 2018 to April 2019, an independent external test cohort of 78 patients (mean age, 46 years; range, 30–74 years) from Hubei Cancer Hospital, Hubei, China, was screened with the same criteria used for the primary cohort. A flowchart describing the research process is shown in Figure 1. Baseline clinicopathologic data, including age, sex, pathologic findings, and US diagnosis reports, were derived from the medical records. US images were obtained from the breast imaging databases at Tongji Hospital and Hubei Cancer Hospital. The clinical US diagnoses were made by 11 radiologists from Tongji Hospital and three radiologists from Hubei Cancer Hospital according to standard protocols (11). For each patient, one or two of the most representative images were selected by three radiologists from Tongji Hospital (L.Q.Z., X.W.C., and G.G.W., each with 5–6 years of experience) for image quality control based on the pathologic results. If the location and size of the lesion on the US image did not match that seen at pathologic examination, we considered removing the case. US equipment manufactured by Philips (Amsterdam, the Netherlands; EPIQ5, EPIQ7 and IU22), Samsung (Seoul, South Korea; RS80A), and GE Healthcare (Pittsburgh, Pa; LOGIQ E9, LOGIQ S7) was used to generate the US images. The clinical characteristics of all patients are shown in Table 1.

Data Preprocessing

Because of the limited amount of training data in our data set, we used data augmentation techniques for image processing (12). Data augmentation can artificially increase the training image data set up to 10 times its original size by means of random geometric image transformations, including flipping, rotation, scaling, and shifting. In addition, it can also ensure that the model used focuses on breast cancer lesions rather than various sources of noise (13). All augmented images were resized to 200 × 300 pixels to standardize the distance scale. The data augmentation strategy has been proven to help prevent network overfitting and memorization of the exact details of the training images (14). All preprocessing steps were conducted in Python (version 3.6.6; Python Software Foundation, Wilmington, Del) by using the Keras ImageDataGenerator (*https://keras.io/preprocessing/image/*).

Deep Neural Network

At present, convolutional neural networks (CNNs) are the most well-known type of deep learning architecture in the field of medical image analysis (15), and they perform well in processing data in multiple arrays (eg, two-dimensional image, three-dimensional video or volumetric image). CNNs consist of an input layer, a hidden layer, and an output layer. The hidden layer generally contains convolutional layers, pooling layers, and fully connected layers. Given their advantages, three representative deep CNN architecture models—namely, Inception V3, Inception-ResNet V2, and ResNet-101, pretrained with ImageNet (*http://www.image-net.org/*)—were evaluated to predict lymph node metastasis based on US images depicting breast cancer. These different pretrained networks were obtained from an open access li-

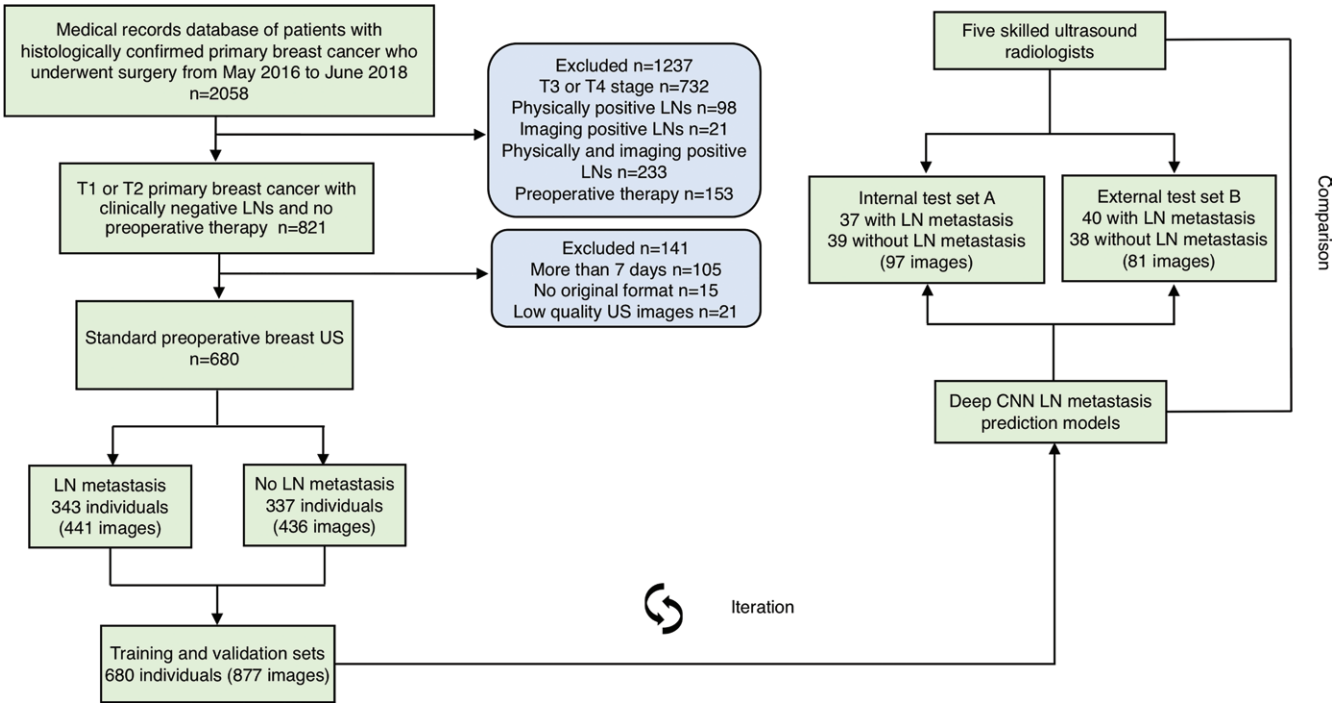

Figure 1: Flowchart of procedures in the development and evaluation of deep learning models for automated lymph node (LN) metastasis prediction. CNN = convolutional neural network.

Table 1: Demographic Data for 834 Patients

Characteristic	Training and Validation Sets	Test Set A	Test Set B
No. of patients	680	76	78
Lymph node metastasis	343 (50)	37 (49)	40 (51)
No lymph node metastasis	337 (50)	39 (51)	38 (49)
Age*	48.6 (24–81)	50.4 (25–82)	45.7 (30–74)
<40 y	112 (17)	12 (16)	9 (12)
40–49 y	265 (39)	30 (40)	32 (41)
50–59 y	185 (27)	21 (28)	25 (32)
60–69 y	90 (13)	10 (13)	10 (13)
≥70 y	25 (4)	3 (4)	2 (3)
Clinical tumor size			
T1 (≤2.0 cm)	326 (47.9)	33 (43.4)	31 (39.7)
T2 (2.1–5.0 cm)	354 (52.1)	43 (56.6)	47 (60.3)
Histologic type			
Ductal	345 (50.7)	36 (47.4)	46 (59.0)
Lobular	214 (31.5)	18 (23.7)	22 (28.2)
Mixed	121 (17.8)	22 (28.9)	10 (12.8)
No. of tumor images	877	97	81
Lymph node metastasis	441 (50.3)	49 (50.5)	43 (53.1)
No lymph node metastasis	436 (49.7)	48 (49.5)	38 (46.9)
Right	454 (51.8)	49 (50.5)	25 (30.9)
Left	423 (48.2)	48 (49.5)	56 (69.1)
Size			
≤2.0 cm	368 (42)	44 (45.4)	36 (44.4)
2.1–4.0 cm	432 (49.3)	38 (39.2)	39 (48.1)
>4.0 cm	77 (8.8)	15 (15.5)	6 (7.4)

Note.—Unless otherwise specified, data in parentheses are percentages.
* Numbers in parentheses are the range.

brary (Keras Applications, available at *https://keras.io/applications/*). More information about the deep CNN models can be found in Appendix E1 (online).

Model Training and Interpretation

All three models were trained by using Keras 2.2.0 with TensorFlow 1.9.0 as the backend. The weight of the network was initialized according to the weight from the pretrained model on ImageNet. The Adam optimizer was used to train the network with a batch size of 32 and adjusted parameters within the CNN by end-to-end supervised learning (16). The initial learning rate was set to 0.0001 and decayed by a factor of 10 each time when there was no further improvement in the accuracy of the validation set for 10 continuous epochs. Finally, the model with the lowest validation loss was selected. During the training phase, the dropout strategy on the fully connected layers with a probability of 0.5 and L2 regularization strategy on weight

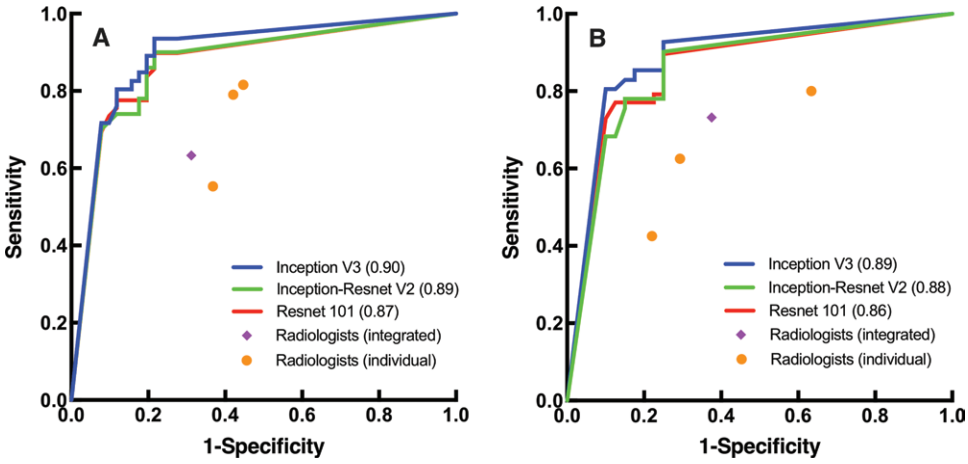

Figure 2: Receiver operating characteristic curves of three deep convolutional neural network models and expert (specificity and sensitivity) points of radiologists for, *A*, test set (10% of data) from Tongji Hospital (test set A) and, *B*, independent external test set from Hubei Cancer Hospital (test set B). *P* values for comparison of Inception V3 versus Inception-ResNet V2, Inception V3 versus ResNet-101, and Inception-ResNet V2 versus ResNet-101 are .44, .33, and .38, respectively, for test set A and .41, .32, and .40 for test set B. Numbers in parentheses are areas under the receiver operating characteristic curves.

and bias were used to prevent the overfitting problem (17). All programs were run in Python version 3.6.6.

To better interpret the network predictions, we used the method of class activation mapping to produce heat maps to visualize the areas of the image most indicative of lymph node metastasis (18). The feature map required to generate the class activation mapping was extracted from the final convolutional layer after the images passed through the fully trained network. All heat maps were produced by using the packages Matplotlib (*https://pypi.org/project/matplotlib/*) and OpenCV (*https://opencv.org/releases.html*).

Clinical Interpretation of US Images

To obtain readers' predictive performance on the independent test set, three board-certified radiologists (Y.B.D., L.Y.B., and S.Y.H., with 26, 14, and 6 years of experience, respectively) performed independent interpretations of the 178 US images from test sets A (n = 97) and B (n = 81). They underwent training in advance in how to perform predictive analysis based on some typical characteristics, such as the size of the primary tumor and the presence of lymphatic invasion, calcifications, and architectural distortion. Interpretations consisted of two components. First, readers performed qualitative evaluation of the primary breast cancer US images with use of the American College of Radiology Breast Imaging Reporting and Data System. Then, they rated the probability of axillary lymph node metastasis (1%–100%) for quantitative prediction analysis. Appendix E1 (online) describes the scoring system and how it is used. Only the primary breast cancer US image data, name, age, and date of examination were visible to the radiologists. The performance of the radiologists was evaluated by comparing their predictions with the pathologic results, which are the diagnostic reference standard. If the predictive results of the three radiologists were inconsistent, the US image was interpreted by two additional breast radiologists (Q.W. and H.R.Y., with 3 and 15 years of experience, respectively). The final prediction determinant of human experts was based on the majority opinion of five radiologists.

Model Testing and Statistical Analysis

Three trained CNN models were tested on two test data sets: 10% of the data from Tongji Hospital were used as the internal hold-out test set A (n = 97), and an independent data set from Hubei Cancer Hospital was used as the external test set B (n = 81). The trained deep learning model outputted the predicted probability of the presence of lymph node metastasis according to US images and chose the class with the highest probability as the prediction result. The prediction scores derived from the deep learning models were compared with the pathology reports. All computer codes used for modeling and data analysis are stored in GitHub (ID: a00d95a; *https://github.com/cakuba/MetastasisPrediction_DeepLearning*). The instructions for using the deep CNN models and data are shown in Appendix E1 (online).

To compare the performances of the deep learning models and radiologists, receiver operating characteristic (ROC) curves were constructed. The specificity and sensitivity points of the radiologists for the independent test sets were plotted in the same ROC space, and areas under the ROC curve (AUCs) with 95% confidence intervals (CIs) were calculated. Comparisons between AUCs were made by using the method devised by DeLong et al (19). The accuracy, sensitivity, and specificity values with 95% CIs were reported for both the deep learning models and radiology readers. κ values and F1 scores were also reported. $P < .05$ was considered to indicate a statistically significant difference. The number of true-positive, false-positive, true-negative, and false-negative findings of the three models and radiologists' performance on test sets were described in a 2×2 contingency table representing the confusion matrix. The statistical analyses were performed by using software (SPSS, version 21.0 [IBM, Armonk, NY], and MedCalc, version 11.2 [MedCalc Software, Ostend, Belgium]). Model training, testing, and visualization were performed by X.L.W; statistical analysis was performed by L.Q.Z.

Results

Clinical-Pathologic Data

As shown in Table 1, we obtained US images depicting primary breast cancer for the training and validation sets (877 images from 680 patients who underwent imaging between May 2016 and June 2018) from the breast imaging database at Tongji

Table 2: Performance of Three CNN Models and Radiologists according to Test Set

Finding	Inception V3	Inception-ResNet V2	ResNet-101	Radiologists	*P* Value
Test set A (*n* = 97)					
Accuracy	80 (78/97) [73, 88]	82 (80/97) [75, 90]	78 (76/97) [70, 87]	66 (64/97) [57, 75]	.99
Sensitivity	82 (40/49) [68, 91]	80 (39/49) [65, 89]	78 (38/49) [63, 88]	63 (31/49) [48, 76]	.14
Specificity	79 (38/48) [65, 89]	85 (41/48) [72, 94]	79 (38/48) [65, 89]	69 (33/48) [54, 81]	.26
PPV	80 (40/50) [66, 90]	85 (39/46) [71, 93]	79 (38/48) [65, 89]	67 (31/46) [52, 80]	.22
NPV	81 (38/47) [66, 90]	80 (41/51) [67, 90]	78 (38/49) [63, 88]	65 (33/51) [50, 77]	.19
κ value	0.61	0.65	0.57	0.32	…
F1 score	0.81	0.82	0.78	0.65	…
Test set B (*n* = 81)					
Accuracy	79 (64/81) [70, 88]	77 (62/81) [67, 86]	73 (59/81) [63, 83]	68 (55/81) [58, 78]	.96
Sensitivity	85 (35/41) [70, 94]	78 (32/41) [62, 89]	73 (30/41) [57, 85]	73 (30/41) [57, 85]	.51
Specificity	73 (29/40) [56, 85]	75 (30/40) [59, 87]	73 (29/40) [56, 85]	63 (25/40) [46, 77]	.62
PPV	76 (35/46) [61, 87]	76 (32/42) [60, 87]	73 (30/41) [57, 85]	67 (30/45) [51, 80]	.72
NPV	81 (29/36) [66, 93]	77 (30/39) [60, 88]	73 (29/40) [56, 85]	69 (25/36) [52, 83]	.58
κ value	0.53	0.58	0.46	0.36	…
F1 score	0.81	0.77	0.73	0.70	…

Note.—Unless otherwise specified, data are percentages, with numbers of images in parentheses and 95% confidence intervals in brackets. CNN = convolutional neural network, NPV = negative predictive value, PPV = positive predictive value.

Table 3: Confusion Matrices for Three CNN Models and Radiologists according to Test Set

	Inception V3 (Truth)		Inception-ResNet V2 (Truth)		ResNet-101 (Truth)		Radiologists (Truth)	
Prediction	Nonmetastasis	Metastasis	Nonmetastasis	Metastasis	Nonmetastasis	Metastasis	Nonmetastasis	Metastasis
Test set A								
Nonmetastasis	38	9	41	10	38	11	33	18
Metastasis	10	40	7	39	10	38	15	31
Test set B								
Nonmetastasis	29	6	30	9	29	11	25	11
Metastasis	11	35	10	32	11	30	15	30

Note.—Data are numbers of images. CNN = convolutional neural network.

Hospital. We obtained images for the internal and external test sets from the breast imaging databases at Tongji Hospital (test set A, 97 images from 76 patients who underwent imaging between June 2018 and October 2018) and Hubei Cancer Hospital (test set B, 81 images from 78 patients who underwent imaging between October 2018 and April 2019). The average patient age was 48.6 years (range, 24–81 years) for the training and validation sets, 50.4 years (range, 25–82 years) for test set A, and 45.7 years (range, 30–74 years) for test set B. Of the 834 patients, 427 (51%) had invasive ductal carcinomas, 254 (30%) had lobular carcinomas, and 153 (18%) had mixed carcinomas. The mean pathologic tumor size was 2.9 cm (range, 0.5–5.0 cm). Of the 834 patients, 390 (47%) had T1 tumors and 444 (53%) had T2 tumors. All 834 patients underwent sentinel lymph node biopsy; positive lymph nodes were found in 420 patients. Among the 420 patients with lymph node metastasis, 230 had micrometastases (≤2 mm) and 190 had macrometastases (>2 mm). Of the 420 patients, 329 (78%) had one or two positive lymph nodes and 91 (22%) had three or more positive lymph nodes. There were 335 lymph node–positive patients who underwent axillary lymph node dissection, including 163 with micrometastasis.

Performance of Deep Learning Models

The deep learning models achieved good performance in predicting lymph node metastasis with the use of the primary breast cancer US images of test set A, with AUCs of 0.90 (95% CI: 0.84, 0.95) for the Inception V3 model, 0.89 (95% CI: 0.83, 0.94) for the Inception-ResNet V2 model, and 0.87 (95% CI: 0.82, 0.93) for the ResNet-101 model (*P* = .44, .33, and .38 for Inception V3 vs Inception-ResNet V2, Inception V3 vs ResNet-101, and Inception-ResNet V2 vs ResNet-101, respectively). For test set B, the AUCs were 0.89 (95% CI: 0.83, 0.95) for the Inception V3 model, 0.88 (95% CI: 0.82, 0.94) for the Inception-ResNet V2 model, and 0.86 (95% CI: 0.77, 0.91) for the ResNet-101 model (*P* = .41, .32, and .40 for Inception V3 vs Inception-ResNet V2, Inception V3 vs ResNet-101, and Inception-ResNet V2 vs ResNet-101, respectively) (Fig 2). Compared with the other two models, the Inception V3 model produced the best re-

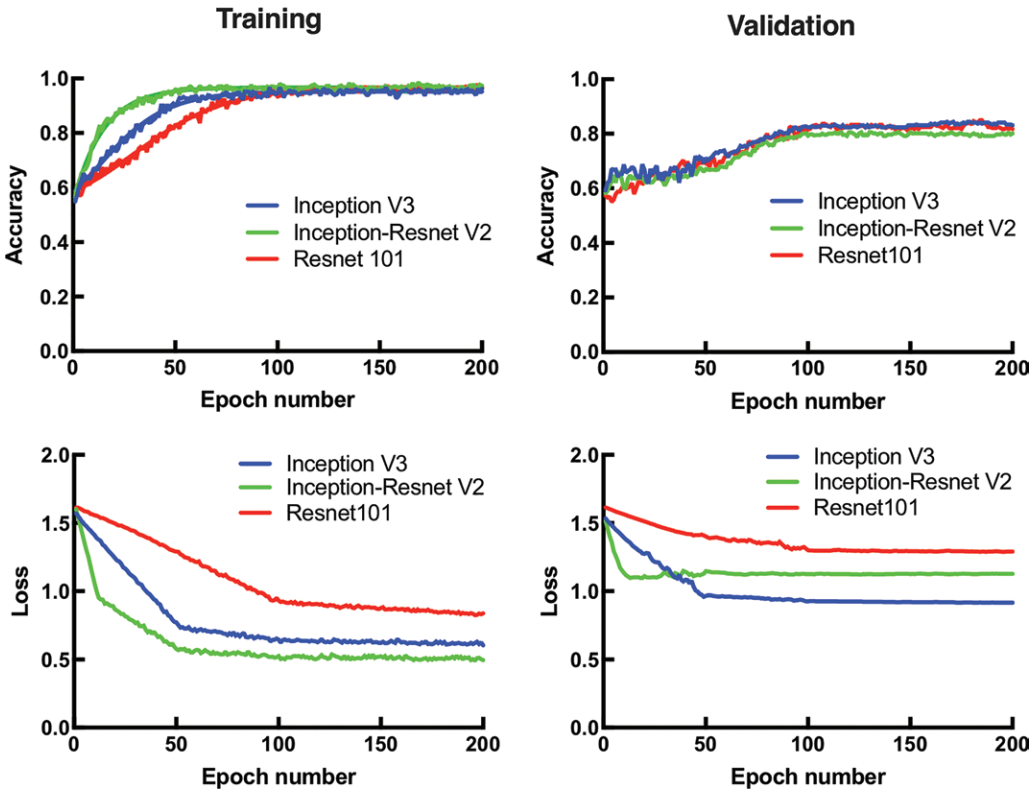

Figure 3: Graphs show training and validation curves of various models. Rising curves represent accuracy of training and validation sets, and falling curves represent loss of training and validation sets, indicating fit between prediction and pathologic truth label.

sults (Table 2). For test set A, the accuracies were 80% (78 of 97 images) for the Inception V3 model, 82% (80 of 97 images) for the Inception-ResNet V2 model, and 78% (76 of 97 images) for the ResNet-101 model (P = .95); the sensitivities were 82% (40 of 49 images; 95% CI: 68%, 91%), 80% (39 of 49 images; 95% CI: 65%, 89%), and 77% (38 of 49 images; 95% CI: 63%, 88%) (P = .88); and the specificities were 79% (38 of 48 images; 95% CI: 65%, 89%), 85% (41 of 48 images; 95% CI: 72%, 94%), and 79% (38 of 48 images; 95% CI: 65%, 89%) (P = .66), respectively. For test set B, the accuracies were 79% (64 of 81 images) for the Inception V3 model, 77% (62 of 81 images) for the Inception-ResNet V2 model, and 73% (59 of 81 images) for the ResNet 101 model (P = .90); the sensitivities were 86% (35 of 41 images; 95% CI: 70%, 94%), 78% (32 of 41 images; 95% CI: 62%, 89%), and 73% (30 of 41 images; 95% CI: 57%, 85%) (P = .40); and the specificities were 73% (29 of 40 images; 95% CI: 56%, 85%), 75% (30 of 40 images; 95% CI: 59%, 87%), and 73% (29 of 40 images; 95% CI: 56%, 85%) (P = .96), respectively. The classification confusion matrices that report the number of true-positive, false-positive, true-negative, and false-negative results for the CNN models and radiology readers are shown in Table 3.

The training curves of the three deep learning classifiers that show the particle process of training are provided in Figure 3. As time goes by, the accuracy of the validation set is represented by the rising solid curves with a final best accuracy of 82.5% at the final 300th epoch, and the loss on the training and validation sets is represented by the falling curves. The similarity of the loss curves of the training and validation sets suggests that the model has no significant overfitting. Training was performed for 300 epochs, and each epoch represents one pass through the entire training set. The final assessment on the test set was performed at epoch 300.

As shown in the heat maps produced by means of the class activation mapping method (Figs 4, 5), the red and yellow regions represent areas activated by the deep CNN model and have the greatest predictive significance; the green and blue backgrounds reflect areas with weaker predictive values. The deeper the feature color is, the greater the possibility of the prediction of lymph node metastasis. This shows that the deep learning network focuses on the most predictive image features of lymph node metastasis.

The performance of the CNN models for predicting lymph node metastasis was also compared with those of skilled breast US radiologists with at least 6 years of experience. As reported in Table 2, the accuracies, sensitivities, and specificities of the reader performance were 66% (64 of 97 images; 95% CI: 57%, 75%), 63% (31 of 49 images; 95% CI: 48%, 76%), and 69% (33 of 48 images; 95% CI: 54%, 81%), respectively, for test set A and 68% (55 of 81 images; 95% CI: 58%, 78%), 73% (30 of 41 images; 95% CI: 57%, 85%), and 63% (25 of 40 images; 95% CI: 46%, 77%) for test set B. The ROC curves for model performance and the points for experts (specificity and sensitivity points for radiologist performance on test sets were plotted in the same ROC space as in Fig 2) show that the points representing the performance of the three radiologists lie below the model ROC curves and outside their 95% CIs. Therefore, the

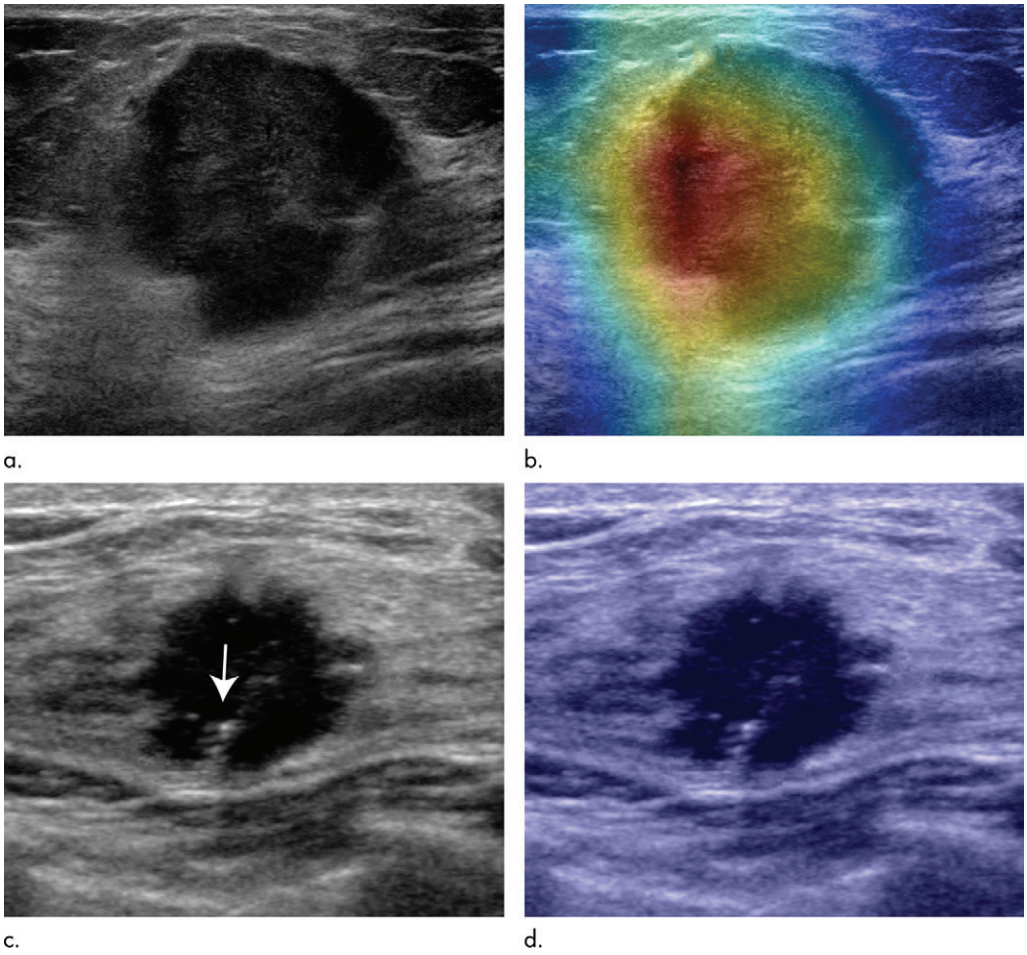

Figure 4: B-mode US images and heat maps of two breast cancers with clinically negative lymph nodes. **(a, b)** Images in a 67-year-old woman with triple-negative invasive ductal and lobular carcinoma and one axillary lymph node metastasis (T2N1). US image (**a**) shows 2.9-cm hypoechoic mass with heterogeneous echotexture. Overlaid heat map (**b**) is an example of true-positive case in which the deep learning model correctly predicted lymph node metastasis, whereas three radiologists did not. **(c, d)** Images in a 46-year-old woman with triple-negative invasive ductal carcinoma and no lymph node metastasis (T1N0). US image (**c**) shows a 1.8-cm irregular hypoechoic mass with multiple internal echogenic foci (arrow). **(d)** True-negative findings were predicted by convolutional neural network models with overlaid heat map (**d**). Three radiologists also correctly predicted no lymph node involvement.

deep learning models performed better than the radiologists in predicting lymph node metastasis on the basis of primary breast cancer US images, with a statistically significant difference.

Discussion

In this study, we successfully developed a predictive model of lymph node metastasis in patients with breast cancer by using deep learning neural networks. The best-performing model yielded satisfactory predictions on the test set, with an area under the receiver operating characteristic curve (AUC) of 0.90, a sensitivity of 82%, and a specificity of 79% for test set A and an AUC of 0.89, a sensitivity of 85%, and a specificity of 72% for test set B. Furthermore, this model significantly outperformed three experienced radiologists in receiver operating characteristic (ROC) space. Our results demonstrate the feasibility of using convolutional neural networks (CNNs) to predict whether early primary breast cancer will metastasize. This work represents an improved approach to the assessment of early lymph node status based on the US images from patients with primary breast cancer obtained before surgery and significantly improves on current prediction methods that rely on physical examinations or lymph node imaging. To the best of our knowledge, this is the first study to apply the deep learning of CNNs for clinically negative lymph node metastasis prediction analysis.

Although US has greatly helped in the direct assessment of lymph node status in patients with breast cancer, such as the features of unclear margins, irregular shapes, or loss of fatty hilum, only visualized nodes can be analyzed and the clinically negative lymph nodes with micrometastasis that have no suspicious imaging features are missed (5). The timely and accurate detection of lymph node micrometastasis is essential for guiding surgical decision making, reconstruction options, and adjuvant therapy. The Memorial Sloan-Kettering Cancer Center and the MD Anderson Cancer Center established a breast cancer nomogram to predict sentinel lymph node positivity by means of a variety of clinical-pathologic factors (20,21), but they did not include other independent predictors, such as imaging features. Multiple studies have shown that the US characteristics of primary breast cancer are

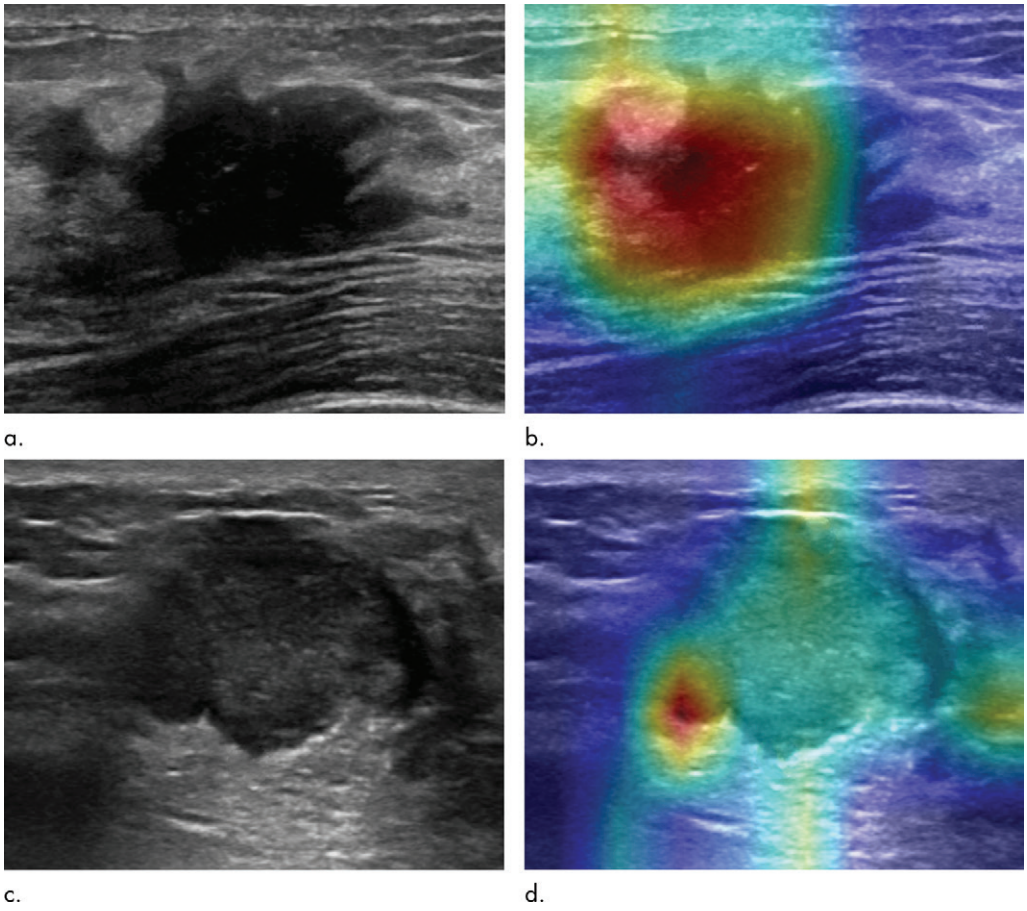

Figure 5: B-mode US images and heat maps of two breast cancers with clinically negative lymph nodes. **(a, b)** Images in a 46-year-old woman with human epidermal growth factor 2–positive ductal carcinoma in situ and no lymph node metastasis (T2N0). US image (**a**) shows a 2.6-cm irregular hypoechoic mass. False-positive findings were predicted by convolutional neural network (CNN) models with overlaid heat map (**b**). Three radiologists correctly predicted no lymph node involvement. **(c, d)** Images in a 36-year-old woman with triple-negative invasive ductal and lobular carcinoma and two axillary lymph node metastases (T1N1). US image (**c**) shows a 2.0-cm hypoechoic mass with heterogeneous echotexture. The CNNs incorrectly predicted no lymph node metastasis on the basis of overlaid heat map (**d**). Two of three radiologists correctly predicted that there was metastasis to axillary lymph nodes.

closely related to axillary lymph node metastasis and have the potential to enable more accurate prediction of the status of clinically negative lymph nodes before surgery (6,22,23). The closer the breast tumor is to the skin and nipple, the greater the likelihood of axillary lymph node metastasis (6,22). If the distance of the tumor from the skin is less than 0.5 cm, the radiologist should thoroughly perform axillary lymph node US and biopsy (24). In addition, the presence of architectural distortions (25), lymphatic invasion (26), and calcifications (23) on breast US images also showed predictive implications for lymph node metastasis. Although studies of these breast features have revealed a great deal about lymph node metastasis, they require human visual assessment based on expertise and experience, which are operator dependent, and qualitative analyses may suffer from data loss owing to individuals with uncertain results.

More recently, deep learning has made substantial progress, allowing machines to automatically represent and explain complicated data (9), and CNN-based image analysis has been applied to establish a direct link between diagnostic images and disease prediction. For example, Ding et al (27) recently showed that a deep learning algorithm–based approach performed better than radiologists in the prediction of Alzheimer disease with fluorine-18 fluorodeoxyglucose PET of the brain. Ha et al (28) demonstrated that a CNN based on a breast MRI tumor data set could predict neoadjuvant chemotherapy response before the initiation of chemotherapy. In our study, we showed that a deep learning model with a CNN-based method was able to predict lymph node metastasis by using US features of primary breast tumors. In contrast to human visual assessment with partial breast US signs, the deep learning algorithm made its final predictions based on the holistic features of breast US images with varying degrees of influence from various anatomic areas. These findings highlight the advantages of the deep learning algorithm that treats the breast as a pixel-by-pixel volume in the task of prediction, and this quantitative assessment of breast imaging information can result in more accurate and reproducible imaging diagnoses than qualitative reasoning.

The uninterpretable neural network system with applications in medical imaging is usually dubbed "black box"

medicine (29). Understanding how the algorithm discriminates input data and establishes links with predictive labels is hard (30). This is an important consideration, as one wants to confirm that the deep CNN model focused on the US features of breast cancer associated with lymph node metastasis rather than the nonrelevant parts of the image. The method of visualization with a heat map can solve this problem by showing the predictive parts of the image. After being given a US image with lymph node metastasis, the deep CNN model will exhibit the strongest activation regions with various colors that correspond to areas with metastatic features. This feature visualization method provides more confidence in the predictive ability of deep neural networks. Examples of false-negative and false-positive findings are shown herein. The irregular hypoechoic features of these two cases may not be consistent with the characteristics associated with lymph node metastasis that the deep learning models learned during training.

To achieve individualized and precise minimally invasive treatment, an increasing number of studies have focused on how to select an axillary management strategy to reduce the use of axillary lymph node dissection for positive sentinel lymph nodes and how to provide an option to avoid sentinel lymph node biopsy for clinically lymph node–negative breast cancer (31). Reliable evidence from the American College of Surgeons Oncology Group Z0011 randomized trial demonstrated that axillary lymph node dissection could be omitted in selected patients with T1 or T2 breast cancer and fewer than two positive sentinel lymph nodes because it would not affect overall and disease-free survival (32). Our deep learning model can achieve this important goal by predicting lymph node metastasis based on a noninvasive inspection with the possibility of screening patients with positive lymph nodes. This further emphasizes the great clinical application value of our deep learning model in adding information predictive of lymph node metastasis.

Our study has several limitations. First, this was a retrospective study, and the results were dependent on the composition of these limited-size data. Further improvement with larger and prospective studies must be achieved before actual clinical use. Second, although all US examinations were performed under the supervision of experienced physicians, there was some variability in the quality of the images because the examinations were performed by multiple physicians. Third, lymph node metastasis and nonmetastasis were inherently unstable diagnoses in that their accuracy is dependent on the time of breast surgery. For example, some of the patients with negative lymph nodes, if followed up for a long enough time, may have eventually progressed to have positive lymph nodes.

In conclusion, we demonstrated that a deep learning algorithm can predict with high accuracy the final diagnosis of lymph node metastasis from two-dimensional gray-scale US images of primary breast cancer. This strategy may be an effective alternative to early screening for lymph node metastasis in clinically lymph node–negative breast cancer. With further validation in a larger population and model calibration, our convolutional neural network–based model has great potential to serve as an important decision support tool in clinical applications.

Acknowledgment: The authors thank Shu-E Zeng, MD, Department of Medical Ultrasound, Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology, for assisting with collection of the imaging data used in this study.

Author contributions: Guarantors of integrity of entire study, L.Q.Z., S.Y.H., G.G.W., Q.W., X.W.C.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, L.Q.Z., S.Y.H., G.G.W., Q.W., Y.B.D., X.R.L., X.W.C., C.F.D.; clinical studies, L.Q.Z., S.Y.H., G.G.W., H.R.Y., Q.W., L.Y.B., Y.B.D., X.R.L., X.W.C.; experimental studies, L.Q.Z., X.L.W., S.Y.H., G.G.W., H.R.Y., Q.W., L.Y.B., X.R.L., X.W.C.; statistical analysis, L.Q.Z., S.Y.H., G.G.W., H.R.Y., Q.W., Y.B.D., X.R.L., X.W.C.; and manuscript editing, L.Q.Z., S.Y.H., G.G.W., H.R.Y., Q.W., Y.B.D., X.R.L., X.W.C., C.F.D.

Disclosures of Conflicts of Interest: **L.Q.Z.** disclosed no relevant relationships. **X.L.W.** disclosed no relevant relationships. **S.Y.H.** disclosed no relevant relationships. **G.G.W.** disclosed no relevant relationships. **H.R.Y.** disclosed no relevant relationships. **Q.W.** disclosed no relevant relationships. **L.Y.B.** disclosed no relevant relationships. **Y.B.D.** disclosed no relevant relationships. **X.R.L.** disclosed no relevant relationships. **X.W.C.** disclosed no relevant relationships. **C.F.D.** disclosed no relevant relationships.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018;68(1):7–30.
2. Tafreshi NK, Kumar V, Morse DL, Gatenby RA. Molecular and functional imaging of breast cancer. Cancer Contr 2010;17(3):143–155.
3. Yenidunya S, Bayrak R, Haltas H. Predictive value of pathological and immunohistochemical parameters for axillary lymph node metastasis in breast carcinoma. Diagn Pathol 2011;6(1):18.
4. Senkus E, Kyriakides S, Ohno S, et al. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2015;26(Suppl 5):v8–v30.
5. de Boer M, van Deurzen CH, van Dijck JA, et al. Micrometastases or isolated tumor cells and the outcome of breast cancer. N Engl J Med 2009;361(7):653–663.
6. Ansari B, Morton MJ, Adamczyk DL, et al. Distance of breast cancer from the skin and nipple impacts axillary nodal metastases. Ann Surg Oncol 2011;18(11):3174–3180.
7. Fujii T, Yajima R, Tatsuki H, et al. Significance of lymphatic invasion combined with size of primary tumor for predicting sentinel lymph node metastasis in patients with breast cancer. Anticancer Res 2015;35(6):3581–3584.
8. Yi A, Moon WK, Cho N, et al. Association of tumour stiffness on sonoelastography with axillary nodal status in T1 breast carcinoma patients. Eur Radiol 2013;23(11):2979–2987.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444.
10. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(8):500–510.
11. American Institute of Ultrasound in Medicine; American Society of Breast Surgeons. AIUM practice guideline for the performance of a breast ultrasound examination. J Ultrasound Med 2009;28(1):105–109.
12. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, Lake Tahoe, Nevada: ACM-DL 2012; 1097–1105. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.
13. Roth HR, Lu L, Liu J, et al. Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE Trans Med Imaging 2016;35(5):1170–1181.
14. Kayalibay B, Jensen G, van der Smagt P. CNN-based segmentation of medical imaging data. CoRR 2017; arXiv:1701.03056. https://arxiv.org/abs/1701.03056. Accessed January 25, 2019.
15. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
16. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. Computer Science 2014; arXiv:1412.6980. https://arxiv.org/abs/1412.6980. Accessed January 23, 2019.
17. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–1958. http://jmlr.org/papers/v15/srivastava14a.html.
18. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep fea- tures for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016; 2921–2929.
19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
20. Memorial Sloan Kettering Cancer Center breast cancer nomogram. http://nomograms.mskcc.org/breast/index.aspx. Accessed June 18, 2019.

21. MD Anderson Cancer Center breast cancer nomogram. http://www3.mdanderson.org/app/medcalc/bc_nomogram3/index.cfm?pagename=sln. Accessed June 19, 2019.
22. Torstenson T, Shah-Khan MG, Hoskin TL, et al. Novel factors to improve prediction of nodal positivity in patients with clinical T1/T2 breast cancers. Ann Surg Oncol 2013;20(10):3286–3293.
23. Bae MS, Shin SU, Song SE, Ryu HS, Han W, Moon WK. Association between US features of primary tumor and axillary lymph node metastasis in patients with clinical T1-T2N0 breast cancer. Acta Radiol 2018;59(4):402–408.
24. Stachs A, Thi AT, Dieterich M, et al. Assessment of ultrasound features predicting axillary nodal metastasis in breast cancer: the impact of cortical thickness. Ultrasound Int Open 2015;1(1):E19–E24.
25. Cho N, Moon WK, Han W, Park IA, Cho J, Noh DY. Preoperative sonographic classification of axillary lymph nodes in patients with breast cancer: node-to-node correlation with surgical histology and sentinel node biopsy results. AJR Am J Roentgenol 2009;193(6):1731–1737.
26. Cornwell LB, McMasters KM, Chagpar AB. The impact of lymphovascular invasion on lymph node status in patients with breast cancer. Am Surg 2011;77(7):874–877.
27. Ding Y, Sohn JH, Kawczynski MG, et al. A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using ^{18}F-FDG PET of the Brain. Radiology 2019;290(2):456–464.
28. Ha R, Chin C, Karcich J, et al. Prior to Initiation of Chemotherapy, Can We Predict Breast Tumor Response? Deep Learning Convolutional Neural Networks Approach Using a Breast MRI Tumor Dataset. J Digit Imaging 2019;32(5):693–701.
29. Ford RA, Price WN. Privacy and accountability in black-box medicine. Social Science Electronic Publishing 2016;23(1):1–43. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2758121. Accessed January 15, 2019.
30. Augasta MG, Kathirvalavakumar T. Reverse engineering the neural networks for rule extraction in classification problems. Neural Process Lett 2012;35(2):131–150.
31. Morrow M. Management of the node-positive axilla in breast cancer in 2017: Selecting the right option. JAMA Oncol 2018;4(2):250–251.
32. Giuliano AE, Ballman KV, McCall L, et al. Effect of Axillary Dissection vs No Axillary Dissection on 10-Year Overall Survival Among Women With Invasive Breast Cancer and Sentinel Node Metastasis: The ACOSOG Z0011 (Alliance) Randomized Clinical Trial. JAMA 2017;318(10):918–926.