Weakly Supervised 3D Deep Learning for Breast Cancer Classification and Localization of the Lesions in MR Images

Juan Zhou, MD,[1] Lu-Yang Luo, MS,[2] Qi Dou, PhD,[2] Hao Chen, PhD,[2,3]* Cheng Chen, MS,[2] Gong-Jie Li, MD,[1]* Ze-Fei Jiang, MD,[4] and Pheng-Ann Heng, PhD[2]

Background: The usefulness of 3D deep learning-based classification of breast cancer and malignancy localization from MRI has been reported. This work can potentially be very useful in the clinical domain and aid radiologists in breast cancer diagnosis.
Purpose: To evaluate the efficacy of 3D deep convolutional neural network (CNN) for diagnosing breast cancer and localizing the lesions at dynamic contrast enhanced (DCE) MRI data in a weakly supervised manner.
Study Type: Retrospective study.
Subjects: A total of 1537 female study cases (mean age 47.5 years ±11.8) were collected from March 2013 to December 2016. All the cases had labels of the pathology results as well as BI-RADS categories assessed by radiologists.
Field Strength/Sequence: 1.5 T dynamic contrast-enhanced MRI.
Assessment: Deep 3D densely connected networks were trained under image-level supervision to automatically classify the images and localize the lesions. The dataset was randomly divided into training (1073), validation (157), and testing (307) subsets.
Statistical Tests: Accuracy, sensitivity, specificity, area under receiver operating characteristic curve (ROC), and the McNemar test for breast cancer classification. Dice similarity for breast cancer localization.
Results: The final algorithm performance for breast cancer diagnosis showed 83.7% (257 out of 307) accuracy (95% confidence interval [CI]: 79.1%, 87.4%), 90.8% (187 out of 206) sensitivity (95% CI: 80.6%, 94.1%), 69.3% (70 out of 101) specificity (95% CI: 59.7%, 77.5%), with the area under the curve ROC of 0.859. The weakly supervised cancer detection showed an overall Dice distance of 0.501 ± 0.274.
Data Conclusion: 3D CNNs demonstrated high accuracy for diagnosing breast cancer. The weakly supervised learning method showed promise for localizing lesions in volumetric radiology images with only image-level labels.
Level of Evidence: 4
Technical Efficacy: Stage 1

J. MAGN. RESON. IMAGING 2019;50:1144–1151.

Breast cancer is the most common malignancy affecting women worldwide,[1] and early diagnosis of breast cancer as well as localizing the lesions are essential for successful treatment planning. While mammography is widely used for early screening of breast cancer in current clinical practice, breast magnetic resonance imaging (MRI) is the imaging modality with the highest sensitivity to diagnose breast cancer.[2] In addition, the MRI scan has also been recommended for screening high-risk populations.[3] Particularly, dynamic contrast-enhanced (DCE) MRI can provide accurate information on the location, size, and volume of the lesions, and has become the first-line method for assessment of tumors.[4,5]

The MRI lexicon in the Breast Imaging-Reporting and Data System (BI-RADS) atlas[6] provides a standardized language

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.26721

Received Jan 2, 2019, Accepted for publication Mar 4, 2019.

*Address reprint requests to: H.C., Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. E-mail: hchen@cse.cuhk.edu.hk or G.-J.L., Department of Radiology, The Fifth Medical Center of Chinese PLA General Hospital, 8 DONGDA St, Fengtai district, Beijing, 100071, Peoples' Republic of China. E-mail:ligj307@163.com
The first two authors contributed equally to this work.

From the [1]Department of Radiology, The Fifth Medical Center of Chinese PLA General Hospital, Beijing, 10071, People's Republic of China; [2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong; [3]Imsight Medical Technology Co., Ltd., Nanshan, Shenzhen, 518000, People's Republic of China; and [4]Department of Breast Cancer, The Fifth Medical Center of Chinese PLA General Hospital, Beijing, China, 100071

Additional supporting information may be found in the online version of this article

that allows radiologists to communicate significant findings. For breast cancer, MRI BI-RADS is used to characterize breast lesions based on the shape, margin, internal enhancement characteristics, as well as nonmass enhancement, with some certain descriptors indicating malignancy, e.g., an irregular mass margin, the segmental distribution, the clustered ring, etc. Clinically, radiologists grade the BI-RADS categories into six levels, from level-1 to level-6, indicating that the patient is negative, benign, probably benign, suspicious, highly suggestive of malignancy, and biopsy-proven malignancy, respectively.[6] However, the BI-RADS category assessment suffers from the limitation of interobserver variance and often subjectively relies on the radiologist's experience.[7]

Automatic methods can help to reduce the inter-observer variance and improve reproducibility. However, it is challenging to automatically identify and localize breast cancer based on images, given that the tumors have variable sizes, shapes, and locations.[8] In addition, there exist many other abnormalities in the images, such as mastitis, granuloma, and adenopathy, which also have atypical appearances but should be differentiated from cancer. In recent years, deep learning has been emerging as a technique for image classification and object localization tasks, by taking advantage of its outstanding feature representation capability.[9] Convolutional neural network (CNN) models have been successfully applied to a wide range of radiology applications, including automatic image classification,[10,11] detection,[12,13] and segmentation[14,15] of lesions. For classification tasks, the CNNs take the raw image data as input and extract the features by a hierarchy of layers to learn discriminative patterns. The models are required to predict likelihood according to the cancerous and noncancerous ground truth given by biopsy or surgery. In other words, they are supervised by the pathological ground truth of the case.

Based on the learned features of the network, we can infer locations for a region of interest (ROI) with high activations, i.e., neuron outputs, by means of a weakly supervised method. Specifically, we used only an image-level label, i.e., whether an MRI was malignant or benign, to accomplish the malignant lesion localizing task. Such methods have been studied for 2D medical images such as chest x-ray,[10] fetal ultrasound,[16] and multilesion computed tomography (CT).[17] Compared with 2D methods, 3D weakly supervised approaches have not been well studied yet.

The objective of this study was to develop 3D deep learning models identifying cancer from noncancer as well as to localize cancers in a weakly supervised manner based on DCE-MRI.

Materials and Methods

This retrospective study was approved by our Institutional Review Board and the requirement for written informed consent was waived.

Dataset

The cohort of the study was obtained by an evaluation of our institutional medical records from March 2013 to December 2016. Inclusion criteria were: 1) Images were scanned under the same MR protocol. 2) The lesion had complete pathology results (biopsy or surgery); except that three lesions were considered benign after 3 years follow-up. 3) Imaging reports had definite BI-RADS category diagnosed by two breast radiologists with 12-year experience and a senior radiologist with 15-year experience, who was consulted in case of disagreement. 4) Lesions were a) solitary in one breast or b) in both breasts with the

TABLE 1. Lesion Types and Numbers

Lesion type	Number	
Malignant	Invasive cancer	903
	Ductal carcinoma in situ	83
	Mucinous adenocarcinoma	10
	Papillary carcinoma	10
	Basal cell carcinoma	7
	Paget's disease	4
	Metaplastic cancer	4
	Malignant phyllodes tumor	3
	Lymphoma	2
	Lobular carcinoma in situ	1
	Interstitial stromal sarcoma	1
	Mixed tubular carcinoma	1
	Myoepithelial carcinoma	1
	Eosinophil infiltration	1
	Medullary carcinoma	1
	Spindle cell tumor with carcinogenesis	1
Benign	Fibroadenomas	233
	Adenosis	102
	Papilloma	62
	Inflammation	57
	Hyperplasia	49
	Phyllodes tumor	6
	Cyst	3
	Duct dilatation	2
	Hamartoma	1
	Great sweat gland metaplasia	1
	Spindle cell tumor	1
	Lipoma	1

TABLE 2. MRI Sequences Protocols

MR sequence	T1 W1 nonfat-suppressed	DCE-MRI
Protocol	t1_fl3d_tra_ nonFatSat	t1_fl3d_tra_ minte_fs_ dynal+6+c
Repetition time	8.70 msec	4.53 msec
Echo time	4.70 msec	1.66 msec
Matrix	896×896	384×384
Slice thickness	1.10 mm	1.10 mm
Bandwidth	350 Hz/Px	380 Hz/Px
Flip angle	20°	15°

same BI-RADS and pathological results. Exclusion criteria were: Normal or typical background parenchyma enhancement (BPE) in bilateral breasts was eliminated. Table 1 lists the details of lesions including the types and the corresponding amount.

Breast MRI was conducted with a 1.5T system (Magnetom Espree Pink; Siemens, Erlangen, Germany), equipped with an eight-channel breast coil. Patients were examined in the prone position, with both breasts positioned in the coil cavity. Conventional plain scans were carried out using the following parameters: axial T1WI 3D non-fatsuppressed (repetition time / echo time [TR/TE], 8.7/4.7 msec; matrix 896 × 896; slice thickness, 1.1 mm).

DCE-MRI used a 3D fat-suppressed volumetric interpolated breath-hold examination sequence before and six times after bolus injection of gadopentetate dimeglumine (0.1 mmol/kg; Magnevist; Bayer, Berlin, Germany) at 2 mL/s followed by flushing with 20-mL physiological (0.9%) saline using an automatic injector. Both breasts were examined for 7.5 min in the axisal plane. Parameters of DCE-MRI were: TR/TE, 4.53/1.66 msec; matrix 384 × 384; slice thickness, 1.1 mm. Images of each phase were subtracted automatically. Detailed protocols and imaging parameters are shown in Table 2.

We used a breast MRI dataset of 1537 female study cases (mean age 47.5 ± 11.8 years). There were 1031 positive cases confirmed with breast cancer. In this study, we randomly divided this dataset into training (1073 cases), validation (157 cases), and testing (307 cases) subsets (Table 3). The deep learning models were then trained with DCE-MRI subtraction images and supervised by pathological labels.

TABLE 3. Distribution of the Collected MRI Breast Dataset in This Study

Case type	Training	Validation	Testing	Subtotal
Malignant	720	105	206	1031
Benign	353	52	101	506
Subtotal	1073	157	307	1537

Training data are for training the network. Validation data are for testing the model after each training step and tuning the hyper-parameters. Testing data are for final performance evaluation.

Data Preprocessing for Breast Segmentation

We used the aforementioned imaging sequence of T_1-weighted nonfat-suppressed MRI to generate 3D masks of the breast area. Specifically, we took 2D slices of the MRI and applied Frangi et al's approach[18] to obtain the breast–air boundary, the pectoralis muscle boundary, as well as the boundaries between breast glands and fat. Next, we employed a series of morphological image processing methods, including thresholding of the filtered slice, connected component analysis, and hole-filling to obtain 2D binary masks for the breast region. Then we stacked the 2D masks of all the slices and obtained a 3D breast segmentation mask of the volumetric data. Next, we employed a 3D Gaussian filter (standard deviation 20) to smooth the breast mask. In this way we obtained the bounding-box that covered the whole breast area. An illustration of the preprocessing procedure is shown in Fig. 1. We cropped the breast region out of all DCE-MRI subtractions using the 3D masks. Then we performed normalization by 1) clipping the image intensities to excluding extreme values; 2) transforming the values into range of 0–1; 3) calculating the overall mean and variance of intensity among all MRIs; and 4) subtracting the mean value from all images and dividing them by variance to be the inputs to our model. Figure 2 shows a typical intensity histogram change of one sample before and after the normalization.

3D Deep Learning Network Architecture

3D DenseNet[19] was utilized as the infrastructure of our deep learning model. The DenseNet is a special architecture of CNN, where shallow layers are densely connected to deeper layers. Specifically, our 3D DenseNet had 37 layers, consisting of an initial stem structure, four densely connected blocks, three transition layers, and finally a classification layer for prediction. In the densely connected block, all the features of the former layers were concatenated as input to the latter layer.

More details about our 3D deep learning network architecture and training strategies are described in the online Appendix.

Malignancy Localization With Weakly Supervised Convolutional Networks

We took advantage of a classification activation map (CAM) method[20] for the weakly supervised cancer localization. Particularly, we connected the Dense Block_3 layer to an additional classification layer to predict the whole image as malignancy or benign. The classifier is started with a 1 × 1 × 1 convolutional layer that generated two feature maps. Next, these two feature maps were input to a global average pooling (GAP) or global max pooling (GMP) layer, which can abstract the feature maps into two activation values. GAP calculates the average value, whereas GMP captures the peak activation in each feature map. In this way the activation map generated before the GP layer could be interpreted as a heatmap that presented the likelihood of tumor lesions across spatial locations. High values in the heatmap revealed a high probability of malignancy. As shown in Fig. 3b, fired positions in the heatmaps indicate the presence of tumor patterns. However, it could only detect the approximate locations of the cancers. To refine the localization result, we applied the Dense Conditional Random Field (DenseCRF).[21]

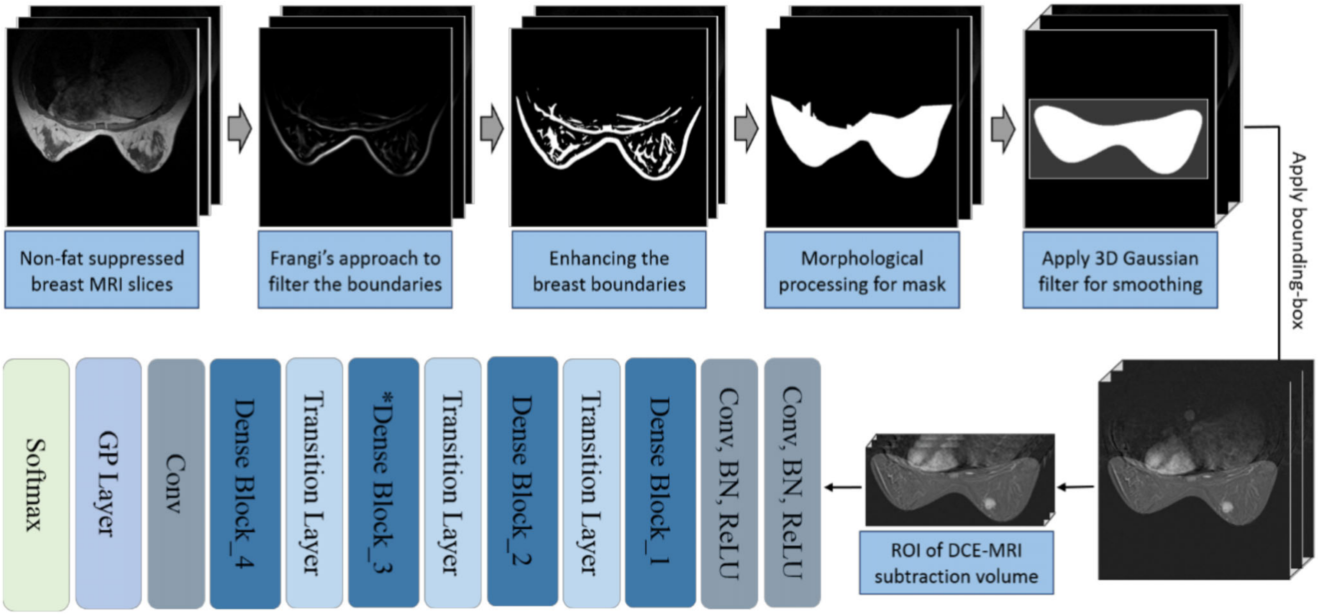
Non-fat suppressed breast MRI slices
Frangi's approach to filter the boundaries
Enhancing the breast boundaries
Morphological processing for mask
Apply 3D Gaussian filter for smoothing
Apply bounding-box
ROI of DCE-MRI subtraction volume
Conv, BN, ReLU
Conv, BN, ReLU
Dense Block_1
Transition Layer
Dense Block_2
Transition Layer
*Dense Block_3
Transition Layer
Dense Block_4
Conv
GP Layer
Softmax

FIGURE 1: Illustration of the framework for breast cancer classification using 3D DenseNet. The preprocessing steps to segment the breast region are presented in the upper part. The convolutional network architecture is presented in the lower part. More details about configurations of the layers are shown in the online Appendix.

To this end, we were able to detect breast cancer in 3D MRI under the setting of weakly supervised learning. The result was both qualitatively and quantitatively evaluated.

Statistical Analysis

The statistical analyses were conducted using MatLab (https://www.mathworks.com/). On testing data, the accuracy, sensitivity, and specificity for diagnosing the breast cancer cases were evaluated. The 95% confidence intervals (CIs) on the metrics were determined using the adjusted Wald method.[22] We obtained the receiver operating characteristic (ROC) curves using the sensitivity and specificity, and calculated the areas under the ROC curves (AUCs) of the models.[23] McNemar's test[24] was used for comparing different models. $P < 0.05$ was considered to indicate statistical significance. The Dice similarity coefficient[25] was used to evaluate the weakly supervised localization method.

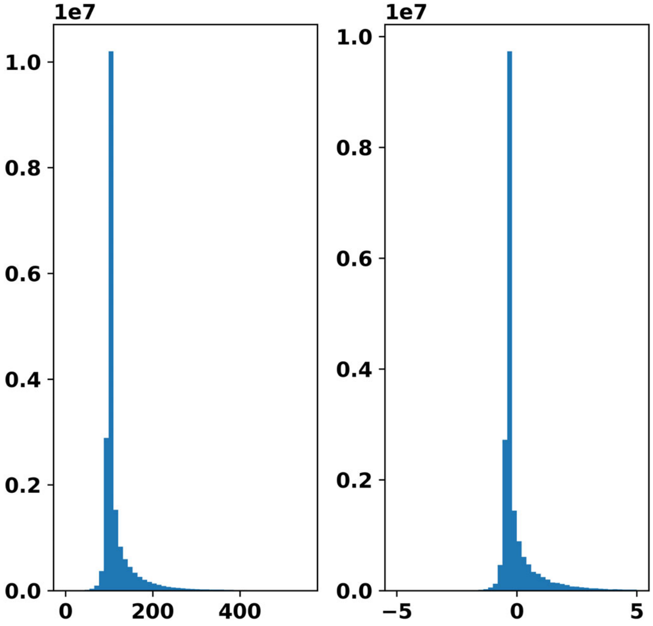

FIGURE 2: Intensity histogram of sample data before and after normalization.

Results

Performance of the CNN Models

We experimented with model configurations of GAP and GMP. The network architectures were identical and their only difference was the operation at the GP layer. Table 4 shows the classification results of the GAP model, the GMP model, and model ensemble. We tested all six DCE-MRI subtractions of one patient and took the average prediction as the final result of one case. By taking 0.5 as the threshold for malignancy prediction, the network conducting GAP obtained an accuracy of 81.1% (95% CI: 76.3%, 85.1%) on the testing subset. By replacing GAP with GMP, the network achieved an accuracy of 81.8% (95% CI: 77.0%, 85.7%). Both networks achieved over 80% accuracy, demonstrating the effectiveness of 3D DenseNets to diagnose breast cancer in MRI scans. Significant differences on the testing accuracies between the two GP settings were not observed ($P = 0.839$).

We also evaluated the sensitivity and specificity of the results produced by the automatic methods. The network using GMP reached a higher sensitivity than the GAP network (91.8% vs. 86.4%, $P < 0.05$). On the other hand, the specificity of the GMP model was lower than that of the GAP network (61.4% vs. 70.3%, $P < 0.05$). This reflects that the GMP made the network more sensitive to the breast cancer tumors, i.e., the GMP network was more aggressive towards malignancy diagnosis.

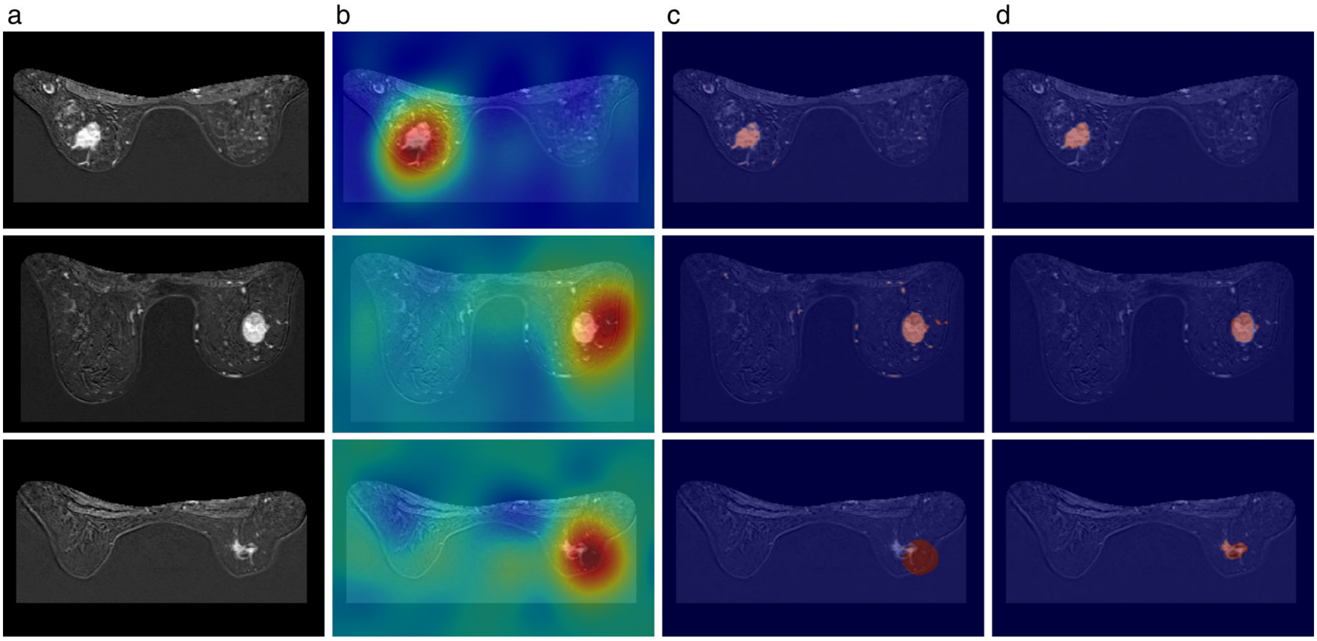

FIGURE 3: Visualization of (a) the MRI slices from three different samples, (b) the corresponding heatmap obtained from the GMP model, (c) the corresponding refined weak label using DenseCRF, and (d) the manual annotation. Fired color indicates higher values for the activations in (b). Red color indicates the annotation by model and human in (c) and (d). The Dice coefficients of each sample were: 0.823, 0.683, and 0.091, respectively.

In addition, we aggregated the outputs of GAP and GMP models by averaging the prediction probabilities of the two DenseNets. As listed in the third row of Table 4, the model ensemble result achieved a higher accuracy (83.7%, 95% CI: 79.1%, 87.4%) compared with the GAP model (P = 0.039) and GMP model (0.146). The balance between sensitivity and specificity was also improved, with a sensitivity of 90.8% at specificity of 69.3%. The AUC of the model ensemble reached 0.859.

For the weakly supervised localization task, one senior radiologist manually annotated 36 correctly predicted cancer samples in the testing data. We calculated the Dice similarity coefficient between the weak labels generated by our method and the human annotations of each subtraction image. The proposed process demonstrated a mean Dice distance of 0.501 and a standard deviation of 0.274.

CNN Performance on Cases With Different BI-RADS Categories

The results for malignancy cases with their BI-RADS categories are shown in Table 5. Our deep learning model reached the highest malignancy sensitivity for cases of Category 5, while it performed worse for the Category 3 cases (92.5% vs. 33.3%,

TABLE 4. Breast Cancer Diagnosis Performance of Deep Learning Networks and Radiologists

Network settings	Accuracy	Sensitivity	Specificity	AUC
3D DenseNet_GAP	81.1% (249/307) [76.3%, 85.1%]	86.4% (178/206) [81.0%, 90.5%]	70.3% (71/101) [60.7%, 78.3%]	0.858
3D DenseNet_GMP	81.8% (251/307) [77.0%, 85.7%]	91.8% (189/206) [87.1%, 94.8%]	61.4% (62/101) [51.6%, 70.3%]	0.856
Model Ensemble	83.7% (257/307) [79.1%, 87.4%]	90.8% (187/206) [86.0%, 94.1%]	69.3% (70/101) [59.7%, 77.5%]	0.859
Radiologist	98.5% (203/206)	59.4% (60/101)	85.7% (263/307)	NA

The accuracy, sensitivity, and specificity are in percentages, with raw data in the parentheses and 95% CIs in brackets. DenseNet = densely connected neural network; GAP = global average pooling; GMP = global max pooling; AUC = area under the receiver operating characteristic curve. BI-RADS 2 and 3 are regarded as benign and 4 and 5 are regarded as malignant diagnosed by radiologists.

TABLE 5. CNN Performance on Malignancy Testing Cases of Different BI-RADS Categories

BI-RADS category	Category 2	Category 3	Category 4	Category 5
Number of cases	0	3	15	188
Number of CNN malignancy prediction	NA	1	13	173
Sensitivity of CNN malignancy prediction	NA	33.3%	86.7%	92.0%

CNN = convolutional neural network; NA = not applicable.

TABLE 6. CNN Performance on Benign Testing Cases of Different BI-RADS Categories

BI-RADS category	Category 2	Category 3	Category 4	Category 5
Number of cases	4	56	31	10
Number of CNN benign prediction	4	42	22	2
Sensitivity of CNN benign prediction	100%	75.0%	71.0%	20.0%

CNN = convolutional neural network.

$P < 0.05$). This indicates that BI-RADS 3 cases are difficult not only for radiologists, but also for deep learning approaches.

Table 6 lists the statistics for benign cases of different BI-RADS categories. Four benign cases were given a BI-RADS Category 2 and our deep neural network correctly diagnosed all of them. The benign sensitivity for Category 3 and 4 cases were 75.0% and 71.0%, respectively. We observed that the predicted benign probabilities for Category 3 cases were significantly higher than those for Category 4 cases (0.66 ± 0.29 vs. 0.47 ± 0.31, $P = 0.0012$). Ten benign cases were misclassified Category 5 by the radiologists. Our model classified these 10 cases with benign probabilities ranging from 0.01 to

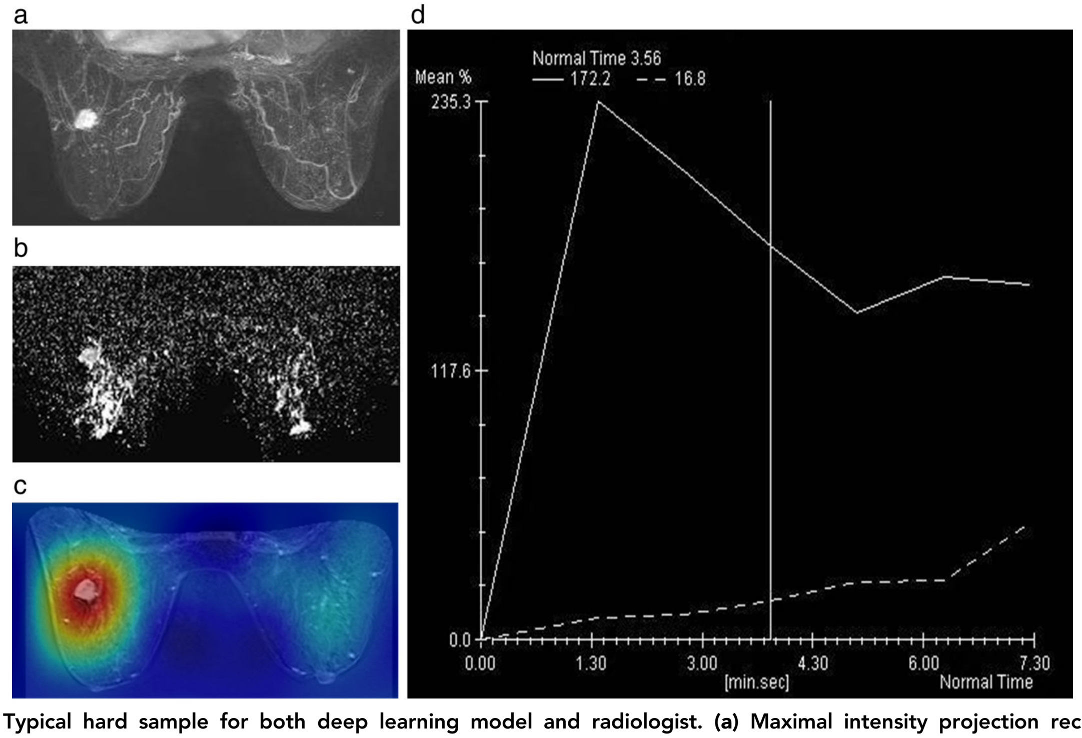

FIGURE 4: Typical hard sample for both deep learning model and radiologist. (a) Maximal intensity projection reconstruction. (b) Apparent diffusion coefficient mapping. (c) Heatmap generated by our model. (d) Time–intensity curve.

0.67, with a mean of 0.17 ± 0.27. Two of the cases received a benign probability of higher than 0.5 and were correctly diagnosed as benign.

Figure 4 shows a typical hard case of intraductal papilloma proved by surgery. Both radiologists and our model identified it as cancer. From the maximal intensity projection reconstruction (Fig. 4a), we can see the irregular shape, irregular margin, and heterogeneous internal enhancement tumor. The time–intensity curve is washout (Fig. 4d). The maximum enhancement is 235.3% within the first 2 minutes after injection. The apparent diffusion coefficient (ADC) mapping (Fig. 4b) suggests water diffusion limited through the tumor tissue (ADC value: 0.78×10^{-3} mm^2/s). Our model showed a high response to this tumor (Fig. 4c) and reported a malignancy probability of 99.8%.

Table 6 compares the performances of CNN with radiologists. By taking BI-RADS 2 and 3 cases as benign, BI-RADS 4 and 5 cases as malignant, classified by radiologists, we observed that the radiologists achieved 98.5% (203 out of 206) sensitivity, 59.4% (60 out of 101) specificity, and 85.7% (263 out of 307) accuracy on the testing subset. CNN improved the specificity while a lower cancer detection rate by nearly 10%. Although CNN could miss more malignancies than the radiologists, it still showed a comparable performance with the radiologists with regard to accuracy.

Visualization of Weakly Supervised Localization

GMP methods generated more compact maps and focused on the most discriminative regions.[20] Hence, we retrieved the heatmaps obtained from the GMP model and utilized DenseCRF to refine the weak annotation. We selected three samples with highest, moderate, and lowest Dice coefficient and show the result in Fig. 3. Typical slices with breast cancer (column (a)) of three cases (each row being one case), slices overlaid by the heatmaps (column (b)), slices overlaid by refined labels (column (c)), and slices overlaid by human annotation (column (d)) are shown in Fig. 3. The three presented cases are malignancy correctly diagnosed by our model.

We can observe that the heatmaps generated from the network are able to locate the cancer areas with higher activations than normal areas (fired color indicates higher activation). These outputs also present an intuitive interpretation of what the models have learned from the training data. The networks were automatically driven to focus on the lesion areas, with the target of making accurate diagnoses. The lesions were considered highly informative and the neuro-activations corresponding to these lesion locations heavily contributed to the final malignancy probability.

Discussion

In this study we directly processed high-dimensional data, i.e., 3D volumetric images, by taking advantage of 3D DenseNet. The layers in our network were densely connected, which encourages reutilization of the features and helps to improve the model performance.[19] In clinical work, breast radiologists give imaging reports by virtue of the image characteristics and their own experience. This strongly depends on reader expertise.[26,27] Moreover, the evaluation of radiologists could be affected by physical fatigue, work environment changes, and many other factors. Regarding the low specificity and high sensitivity of the clinical MRI report, our proposed method missed about 8% of malignancies, while it lowered the overdiagnosing rate by nearly 10%. Therefore, it could serve as an assisting tool in the report system to help raise specificity in cancer screening.

Compared with traditional methods, the main advantage of deep learning models is its capability in extracting highly representative features in a data-driven way. Recently, the efficacy of deep neural networks has been evaluated in breast cancer classification tasks.[28–30] However, these works either used a small-size dataset or needed manual annotations on lesions during the training phase. Directly localizing breast cancers in 3D radiology images with only image-level supervision has not yet been extensively explored. Our work included 1537 cases with pathology labels. Since all the DCE-MRI subtractions are of the same modality, each subtraction can be treated as a training sample by the means of data augmentation. Hence, overall, 9222 sample scans were used to conduct this study. Although using only the structural information in DCE-MRI, the large dataset as well as the informative high-dimensional 3D data enabled greater performance in our study.

The classification performance of all three models suggested that both global and discriminative features in the MR are essential for CNN to identify breast cancer. The comparison between GAP, GMP, and ensemble models indicated that discriminative features contribute the most. In other words, the CNN can learn to focus on the most informative part when analyzing lesions. The results also showed that MRI is highly sensitive for both radiologists and CNNs. It shows that CNNs share some common characteristics with radiologists in identifying breast cancer. In addition, traditionally difficult classified cases were hard for the CNNs as well. The model was less accurate on cases of BI-RADS 3, 4, and benign cases of BI-RADS 5.

Some limitations of this study must be addressed. First, our dataset had a selective nature. Only MRIs with a solitary lesion were included and cases where BPE is evident. We deemed it confusing for the CNN when multiple lesions appeared and asymmetric BPE can be mistaken for nonmass enhancement. Hence, the CNN was trained on relatively easy samples. Second, the network was not trained for identifying specific abnormalities. The network should not only correctly classify a patient as benign, but also distinguish a benign lesion to be mass or mastitis. Consequently, the weakly supervised localization task could only detect the lesions with a high malignancy probability. Lastly, we took each subtraction image as one sample, which meant only structural features were considered in this study.

In conclusion, we developed a 3D deep learning model for breast DCE-MRI cancer classification based on the state-of-the-art densely connection structure. Our deep learning networks demonstrated comparable accuracy with radiologists. The weakly supervised learning method showed promise for localizing lesions in volumetric radiology images with only image-level labels. Our image analyzing pipeline is fully automatic, from breast MRI preprocessing to malignancy likelihood prediction and cancer annotating. The pipeline has potential as an assisting tool under clinical conditions.

Acknowledgments

This project is supported by the National Basic Research Program of China, 973 Program (2015CB351706), the National Natural Science Foundation of China (No.61233012) and Hong Kong Research Grants Council - CUHIK 14225616.

References

1. DeSantis CE, Ma J, Goding Sauer A, Newman LA, Jemal A. Breast cancer statistics, 2017, racial disparity in mortality by state. CA Cancer J Clin 2017;67:439–448.

2. Chiarelli AM, Prummel MV, Muradali D, et al. Effectiveness of screening with annual magnetic resonance imaging and mammography: Results of the initial screen from the Ontario High Risk Breast Screening Program. J Clin Oncol 2014;32:2224–2230.

3. Kuhl C, Weigel S, Schrading S, et al. Prospective multicenter cohort study to refine management recommendations for women at elevated familial risk of breast cancer: The EVA trial. J Clin Oncol 2010;28: 1450–1457.

4. Eisen A, Fletcher GG, Gandhi S, et al. Optimal systemic therapy for early breast cancer in women: A clinical practice guideline. Curr Oncol 2015;22:S67–S81.

5. Hylton NM, Blume JD, Bernreuter WK, et al. Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy—Results from ACRIN 6657/I-SPY TRIAL. Radiology 2012;263:663–672.

6. Rao AA, Feneis J, Lalonde C, Ojeda-Fournier H. A pictorial review of changes in the BI-RADS, 5th ed. Radiographics 2016;36:623–639.

7. Masroor I, Rasoor M, Saeed SA, Sohail S. To asses inter- and intra-observer variability for breast density and BIRADS assessment categories in mammographic reporting. JPMA J Pakistan Med Assoc 2016;66: 194–197.

8. Preim U, Glaser S, Preim B, Fischbach F, Ricke J. Computer-aided diagnosis in breast DCE-MRI—Quantification of the heterogeneity of breast lesions. Eur J Radiol 2012;81:1532–1538.

9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521: 436–444.

10. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017;2097–2106.

11. Dhungel N, Carneiro G, Bradley AP. The automated learning of deep features for breast mass classification from mammograms. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention 2016;106–114.

12. Setio AAA, Traverso A, De Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal 2017;42:1–13.

13. Liu JM, Wang D, Lu L, et al. Detection and diagnosis of colitis on computed tomography using deep convolutional neural networks. Med Phys 2017;44:4630–4642.

14. Kamnitsas K, Ledig C, Newcombe VF, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal 2017;36:61–78.

15. Dou Q, Yu L, Chen H, et al. 3D deeply supervised network for automated segmentation of volumetric medical images. Med Image Anal 2017;41:40–54.

16. Gao Y, Alison Noble J. Detection and characterization of the fetal heartbeat in free-hand ultrasound sweeps with weakly-supervised two-streams convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham, Switzerland: Springer; 2017.

17. Cai J, Tang Y, Lu, et al. Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: Slice-propagated 3D mask generation from 2D RECIST. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham, Switzerland: Springer; 2018. pp 396–404.

18. Frangi AF, Niessen WJ, Vincken KL, Viergever MA. Multiscale vessel enhancement filtering. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham, Switzerland: Springer; 1998. pp 130–137.

19. Huang G, Liu Z, Weinberger KQ, Maaten LVD. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Cham, Switzerland: Springer; 2017. pp 4700–4708.

20. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. Learning deep features for scene recognition using places database. In: *Advances in neural information processing systems.* Cambridge, MA: MIT Press; 2014. pp 487–495.

21. Krähenbühl P, Vladlen K. Efficient inference in fully connected CRFs with Gaussian edge potentials. In: *Advances in neural information processing systems.* Cambridge, MA: MIT Press; 2011. pp 109–117.

22. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat 1998;52:119–126.

23. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003;229:3–8.

24. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12: 153–157.

25. Dice, Lee R. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.

26. Mann RM, Kuhl CK, Kinkel K, Boetes C. Breast MRI: Guidelines from the European Society of Breast Imaging. Eur Radiol 2008;18: 1307–1318.

27. Peters NHGM, Borel Rinkes IHM, Zuithoff NPA, et al. Meta-analysis of MR imaging in the diagnosis of breast lesions. Radiology 2008;246: 116–124.

28. Amit G, Ben-Ari R, Hadad O, Monovich E, Granot N, Hashoul S. Classification of breast MRI lesions using small-size training sets: Comparison of deep learning approaches. In: *Medical Imaging 2017: Computer-Aided Diagnosis.* International Society for Optics and Photonics. 2017. p 101341H.

29. Maicas G, Carneiro G, Bradley AP, Nascimento JC, Reid I. Deep reinforcement learning for active breast lesion detection from DCE-MRI. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Cham, Switzerland: Springer; 2017. pp 665–673.

30. Bickelhaupt S, Paech D, Kickingereder P, et al. Prediction of malignancy by a radiomic signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography. J Magn Reson Imaging 2017;46:604–616.