Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms

Hui Li
Maryellen L. Giger
Benjamin Q. Huynh
Natalia O. Antropova

SPIE.

Hui Li, Maryellen L. Giger, Benjamin Q. Huynh, Natalia O. Antropova, "Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms," *J. Med. Imag.* **4**(4), 041304 (2017), doi: 10.1117/1.JMI.4.4.041304.

Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms

Hui Li,* Maryellen L. Giger, Benjamin Q. Huynh, and Natalia O. Antropova
University of Chicago, Department of Radiology, Chicago, Illinois, United States

Abstract. To evaluate deep learning in the assessment of breast cancer risk in which convolutional neural networks (CNNs) with transfer learning are used to extract parenchymal characteristics directly from full-field digital mammographic (FFDM) images instead of using computerized radiographic texture analysis (RTA), 456 clinical FFDM cases were included: a "high-risk" BRCA1/2 gene-mutation carriers dataset (53 cases), a "high-risk" unilateral cancer patients dataset (75 cases), and a "low-risk dataset" (328 cases). Deep learning was compared to the use of features from RTA, as well as to a combination of both in the task of distinguishing between high- and low-risk subjects. Similar classification performances were obtained using CNN [area under the curve $(AUC) = 0.83$; standard error $(SE) = 0.03$] and RTA ($AUC = 0.82$; $SE = 0.03$) in distinguishing BRCA1/2 carriers and low-risk women. However, in distinguishing unilateral cancer patients and low-risk women, performance was significantly greater with CNN ($AUC = 0.82$; $SE = 0.03$) compared to RTA ($AUC = 0.73$; $SE = 0.03$). Fusion classifiers performed significantly better than the RTA-alone classifiers with AUC values of 0.86 and 0.84 in differentiating BRCA1/2 carriers from low-risk women and unilateral cancer patients from low-risk women, respectively. In conclusion, deep learning extracted parenchymal characteristics from FFDMs performed as well as, or better than, conventional texture analysis in the task of distinguishing between cancer risk populations. [DOI: 10.1117/1.JMI.4.4.041304]

Keywords: deep learning; convolutional neural network; transfer learning; radiographic texture analysis; breast cancer risk assessment; mammographic parenchymal patterns; full-field digital mammogram.

Paper 17029SSRR received Feb. 7, 2017; accepted for publication Aug. 18, 2017; published online Sep. 13, 2017.

1 Introduction

Approximately one in eight women will develop breast cancer in the United States during their lifetime.[1] Breast cancer is the second leading cause of cancer death among women.[2] Currently, mammography continues to be the most effective screening tool for breast cancer early detection.[3]

There are many risk factors for developing breast cancer, including having a family history of breast cancer and inherited genes, such as BRCA1/2 gene mutations, as well as other mutated genes.[4] Personal risk factors include age, personal history of breast cancer, dense breasts, age at menarche, race, radiation exposure, and hormonal therapy. The relationship between mammographic parenchymal patterns and breast density and the risk of developing breast cancer have been studied extensively by using both visual and computerized assessment methods.[5–17] Results from these studies indicated that the breast density is strongly associated with increased risk of developing breast cancer.[8] Overall, women who are at high risk of developing breast cancer tend to have dense breasts, and their mammographic parenchymal patterns tend to be coarse and low in contrast.[10,15]

Recently, the application of deep learning in imaging has been rapidly growing.[18] Deep learning with convolutional neural networks (CNNs) has proved to be a powerful technique in general object recognition, learning high-level image features directly from images, and yielding improved image classification performance.[19] Over the decades, such successes with deep learning have also garnered the interest of researchers in medical imaging analysis.[20–22] However, the training of deep CNNs from scratch is a challenging task, especially in the medical imaging field, since such training requires large medical imaging datasets with necessary human-delineated annotations, which have proven to be difficult and time-consuming to collect. However, a learning technique called "transfer learning" has emerged and is being applied in medical imaging analysis.[23–28] In these situations, pretrained CNNs modeled with either a nonmedical image dataset or medical image dataset from a different modality are applied to clinical decision-making tasks with a relatively small medical imaging dataset. Outputs extracted from layers of the network can serve as features for various medical tasks. For example, in the work of Samala et al.,[26] pretrained deep CNNs modeled on mammograms were used for breast mass lesion detection on digital breast tomosynthesis images. Huynh et al.[28] applied transfer learning with deep CNNs on digital mammograms for the diagnostic classification of breast tumors, with results demonstrating performance levels as current computer-aided diagnosis (CADx) methods.

The purpose of this study is to evaluate the potential of deep learning in the assessment of breast cancer risk, in which CNNs extract parenchymal pattern features directly from full-field digital mammographic (FFDM) images. The classification performance based on CNN-extracted features is compared with that based on texture features extracted from conventional

*Address all correspondence to: Hui Li, E-mail: huili@uchicago.edu

2329-4302/2017/$25.00 © 2017 SPIE

computerized radiographic texture analysis (RTA).[15,17] To the best of our knowledge, this is the first effort to use deep learning to (1) employ deep CNN with transfer learning in breast cancer risk assessment and (2) compare it with conventional texture analysis.

2 Materials and Methods

2.1 *Dataset*

Mammographic cases were retrospectively collected at the University of Chicago Medical Center under an institutional review board approved protocol with a waiver of consent and compliance to the Health Insurance Portability and Accountability Act.

The FFDM images used in this study were acquired with a GE senographe 2000D system (Waukesha, Wisconsin), with 100-μm pixel size and 12-bit quantization. Regions of interest (ROIs) of 256×256 pixels were manually selected from the center breast region behind the nipple (Fig. 1). Only the craniocaudal view of mammographic images was used in the study. Details regarding ROI selection can be found elsewhere.[29]

Four hundred and fifty six cases were included in this study, which included two high-risk datasets and one low-risk dataset as the control group.[17] One high-risk dataset consisted of FFDMs of women with the BRCA1/2 gene-mutation. There were 53 mutation carriers—36 with the BRCA1 gene-mutation and 17 with the BRCA2 gene-mutation. The average age was 40.2 years with a standard deviation of 11.8 years. The other high-risk dataset included 75 women with unilateral breast cancer—43 invasive ductal carcinoma, 7 invasive lobular carcinoma, 14 ductal carcinoma *in situ*, 1 lobular carcinoma *in situ*, 2 Paget's disease, and 8 other atypical malignant diseases. The average age was 55.8 years old with a standard derivation of 15.0 years. The low-risk group included 328 women who could be considered at usual risk and were undergoing screening mammography between 2006 and 2008 at the University of Chicago Medical Center. These women each had a lifetime risk of <10% based on the Gail breast cancer risk assessment model.[30] The average age was 58.4 years with a standard derivation of 11.9 years. The baseline characteristics of the study populations are shown in Table 1.

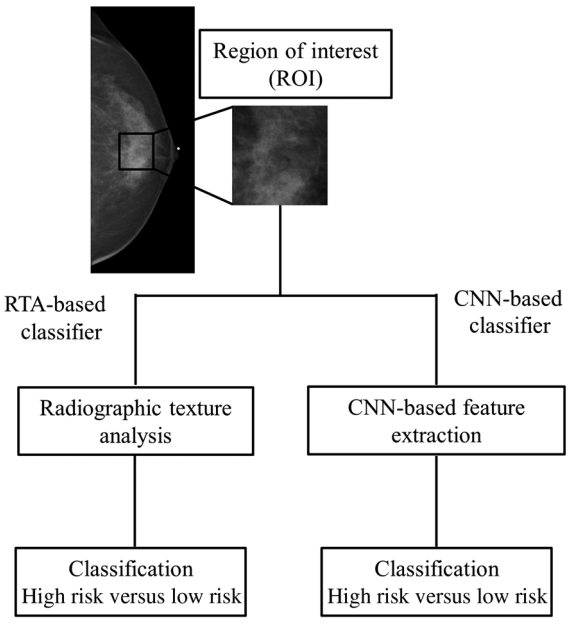

Fig. 1 Schematic diagram of conventional RTA- and deep CNN-based methods for breast cancer risk assessment.

Mammograms used in this study were reviewed by an expert breast mammographer to ensure that there were no detectable abnormalities on the images. As for the unilateral breast cancer women, only their normal contralateral breasts were included in the analysis.

2.2 *Computerized Radiographic Texture Analysis*

Computerized RTA was performed on each individual ROI for texture feature extraction. These texture features were calculated based on gray-level histogram analysis, neighborhood gray-tone difference matrix,[31] gray-level co-occurrence matrix,[32] fractal analysis, edge frequency analysis, and Fourier analysis to characterize the mammographic parenchymal patterns.[33–35] The computer-extracted texture features served as image-based phenotypes to assess the image contrast, image coarseness, image heterogeneity, as well as image local composition, which was related to local density measure. A total of 45 texture features were extracted from each ROI for use in subsequent analyses. Detailed descriptions of these texture features were described elsewhere.[10,15–17,33–37] The RTA-extracted texture features were standardized with zero mean and unit variance prior to input to the classifier.

2.3 *CNN-Based Feature Extraction*

ROI images were used as input to a pretrained CNN to extract CNN-based features.[19] This CNN, i.e., the AlexNet, had been trained on the ImageNet dataset of 1.2 million high-resolution images and used to classify general objects into 1000 classes.[19] The architecture of this pretrained CNN contained five convolutional layers, three pooling layers, and three fully connected layers.[19] Given that the CNN was pretrained, our use of it was restricted to its original architecture and input image size of 227×227 pixels, and thus, 227×227 patches were extracted from the center of each 256×256 ROI. The output from the first fully connected layer, a vector of 4096 in length, served as the CNN-based features, which subsequently underwent dimension reduction by eliminating those features with zero-variance features across the datasets. The CNN-based features were then standardized with zero mean and unit variance prior to input to the classifier. The feature extractions were performed on a computer running openSUSE Linux operating system with 6-core/12-thread Intel Xeon CPU E5-2620 2.10 GHZ and 24GB memory.

2.4 *Classification Based on RTA- and CNN-Based Features*

Two risk-based medical classification tasks were performed in this study. One task was distinguishing BRCA1/2 gene-mutation carriers from the low-risk control group, and the other task was distinguishing unilateral cancer patients from the low-risk control group.

Stepwise feature selection was performed on features extracted from the pretrained CNN method and from the conventional RTA method. The p-value of 0.05 was used for addition and removal of the features in the stepwise feature selection step.[38] The selected features were used as input to linear support

Table 1 Baseline characteristics of study population.

	BRCA1/2 gene-mutation carriers versus low-risk women ($n = 381$)			Unilateral cancer women versus low-risk women ($n = 403$)		
Variable	BRCA1/2 gene-mutation carriers ($n = 53$)	Low-risk women ($n = 328$)	p-value for mutation carriers versus low-risk: p-value from t-test	Unilateral cancer women ($n = 75$)	Low-risk women ($n = 328$)	p-value for unilateral cancer versus low-risk: p-value from t-test
Mean age (SD)	40.2 (11.8)	58.4 (11.9)	<0.0001	55.8 (15.0)	58.4 (11.9)	0.1062
Breast mean percent density (%)	27.2 (18.2)	18.7 (17.4)	0.0013	22.5 (18.4)	18.7 (17.4)	0.0910
			p-value from chi-squared test			p-value from chi-squared test
Race						
White, non-Hispanic	49	107	<0.0001	28	107	0.063
Black, non-Hispanic	3	194		34	194	
Asian	0	7		3	7	
American Indian or Alaskan native	0	1		0	1	
Hispanic	1	8		1	8	
Other/mixed	0	11		9	11	
BI-RADS density rating						
A	4 (7.5%)	34 (10.4%)	<0.0001	5 (6.7%)	34 (10.4%)	0.0452
B	18 (34.0%)	200 (61.0%)		43 (57.3%)	200 (61.0%)	
C	25 (47.2%)	90 (27.4%)		22 (29.3%)	90 (27.4%)	
D	6 (11.3%)	4 (1.2%)		5 (6.7%)	4 (1.2%)	

vector machine (SVM) classifiers for the two classification tasks in an iterated leave-one-case-out cross-validation analyses. Given the moderate dataset size, for the pretrained CNN method, only the top 20 features based on the area under the curve (AUC) values from receiver operating characteristic (ROC) analysis were merged with an SVM classifier.[39,40]

In addition, a fusion method was employed in which the classifier outputs from the RTA-based method and the pretrained

Table 2 Classification performances for the conventional RTA method, CNN-based method, and fusion classifier in the task of breast cancer risk assessment (BRCA1/2 versus low risk; unilateral cancer versus low risk) on FFDM (AUC, area under the curve; SE, standard error; CNN, convolutional neutral network). Bonferroni corrections implemented given the multiple comparisons.

Classification task	Classification method	AUC (SE)	p-value for ΔAUC (significance level) (95% confidence interval)			
BRCA1/2 (53) versus low risk (328)	Conventional RTA method	0.82 (0.03)	0.6706 (0.05) [−0.0856, 0.0551]		0.0089 (0.0167) [−0.0806, −0.0116]	
	CNN-based method	0.83 (0.03)				
			0.3227 (0.025) [−0.0665, 0.0219]			
	Fusion classifier	0.86 (0.03)				
Unilateral cancer (75) versus low risk (328)	Conventional RTA method	0.73 (0.03)	0.0090 (0.025) [−0.1653, −0.0236]		< 0.0001 (0.0167) [−0.1527, −0.0639]	
	CNN-based method	0.82 (0.03)				
			0.3627 (0.05) [−0.0490, 0.0179]			
	Fusion classifier	0.84 (0.02)				

CNN-based method were averaged in order to yield a combined output related to the likelihood of being in a high-risk group.

2.5 *Performance Evaluation and Statistical Analysis*

The classification performance, in terms of distinguishing between the high- and low-risk groups, of each classifier was assessed using ROC analysis with the AUC as the figure of merit.[40] The statistical significance for the difference between the classifiers' performance was evaluated using the ROCKIT software.[41] The Holm–Bonferroni method was applied to correct for multiple comparisons.[42] The kappa coefficient[43] was calculated to measure the agreement between the outputs from the classifier based on RTA features and the outputs from the classifier based on CNN-extracted features.

3 Results

The computational time for the CNN-based feature extraction was ~0.1 s per ROI, and the computational time for the conventional RTA-based feature extraction was ~1 s per ROI.

For the RTA method, features selected from stepwise feature selection step (six features for BRCA1/2 versus low risk; seven features for unilateral cancer versus low risk) were merged using an SVM classifier. From the analyses of the parenchymal patterns, similar classification performance levels were obtained using features extracted with the pretrained CNNs (AUC = 0.83; SE = 0.03) and when using features extracted with the RTA method (AUC = 0.82; SE = 0.03) in the task of distinguishing between BRCA1/2 gene-mutation carriers and the low-risk women (Table 2 and Fig. 2). However, in the task of distinguishing between unilateral cancer patients and the low-risk women, classification performance was significantly higher with the CNN-based method (AUC = 0.82; SE = 0.03) as compared to the RTA method (AUC = 0.73; SE = 0.03) with a p-value of 0.009 (Table 2 and Fig. 3).

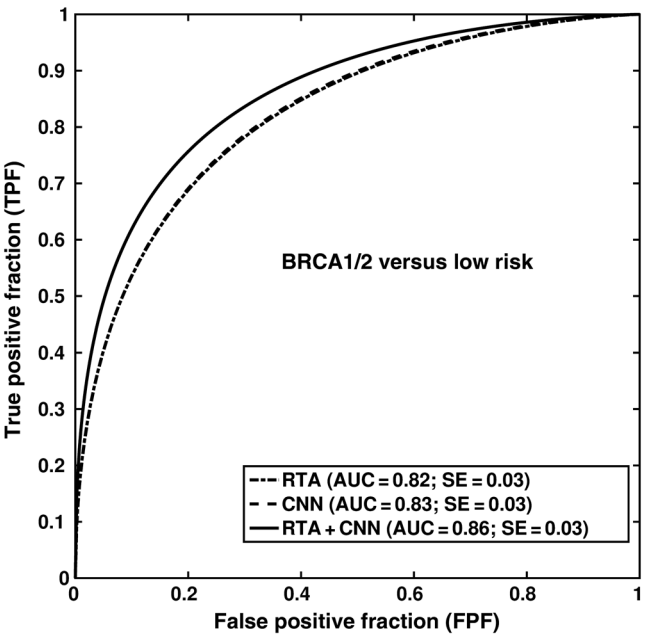

Fig. 2 ROC curves indicating the performance of RTA-based, CNN-based, and fusion classifiers in the task of distinguishing between BRCA1/2 gene-mutation carriers and low-risk women.

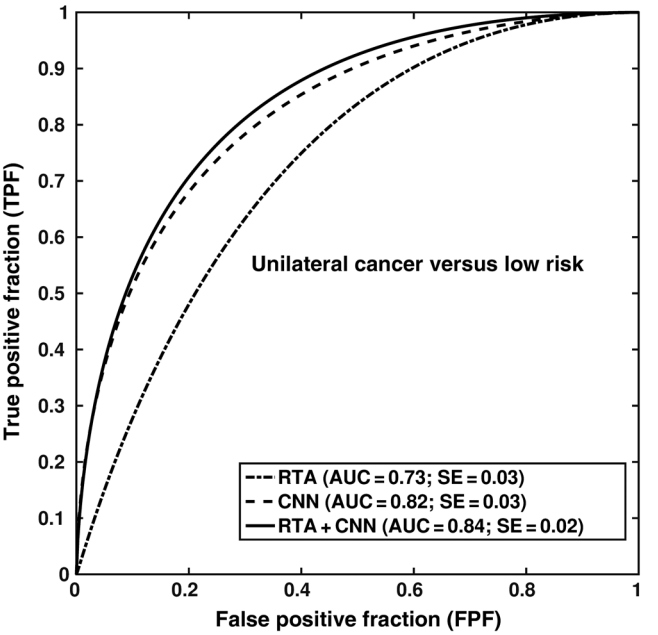

Fig. 3 ROC curves indicating the performance of RTA-based, CNN-based, and fusion classifiers in the task of distinguishing between unilateral cancer patients and low-risk women.

Fair and slight kappa correlations were observed between the classifier outputs using CNN-based features and conventional RTA features in the tasks of distinguishing between BRCA1/2 gene-mutation carriers and the low-risk women (kappa coefficient = 0.2567; $p < 0.0001$; Fig. 4) and between unilateral cancer patients and the low-risk women (kappa coefficient = 0.1682; $p < 0.0044$; Fig. 5). AUC values

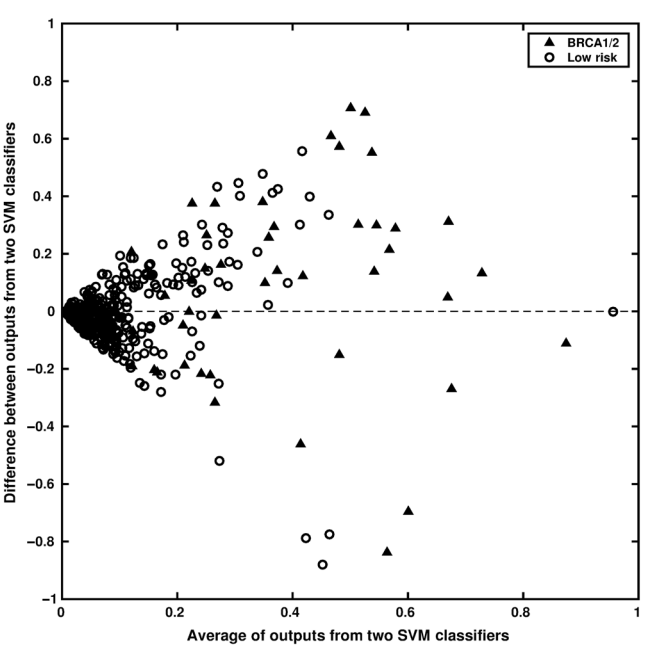

Fig. 4 Bland–Altman plot showing the agreement between the classifier output based on RTA features and the classifier output based on CNN features in the task of distinguishing between BRCA1/2 gene-mutation carriers and low-risk women. The kappa coefficient is 0.2567 (p-value < 0.0001) between the classifier output based on RTA features and the classifier output based on CNN features.

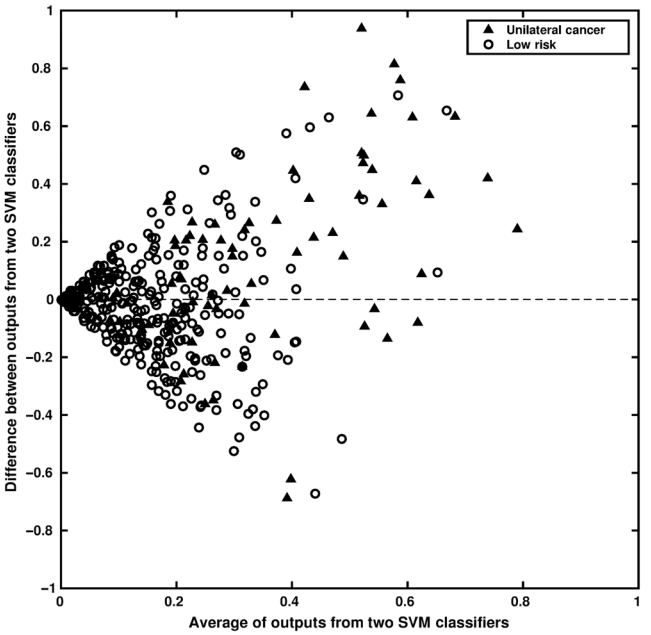

Fig. 5 Bland–Altman plot showing the agreement between the classifier output based on RTA features and the classifier output based on CNN features in the task of distinguishing between unilateral cancer patients and low-risk group. The kappa coefficient is 0.1682 (p-value = 0.0044) between the classifier output based on RTA features and the classifier output based on CNN features.

of 0.86 and 0.84 were obtained for the fusion classifiers in the tasks of differentiating BRCA1/2 gene-mutation carriers from the low-risk group and unilateral cancer patients from the low-risk group, respectively (Table 2). The fusion classifiers performed significantly better than the classifiers based solely on the RTA features ($p = 0.0089$ for BRCA1/2 gene-mutation carriers versus low-risk women; $p < 0.0001$ for unilateral cancer patients versus low-risk women).

The net reclassification improvement between the RTA-based method and the fusion classifier was indicated by the AUC difference and the number of cases and controls correctly reclassified (for BRCA1/2 gene-mutation carriers versus low-risk women, $\Delta\text{AUC} = 0.04$; 2 gene-mutation carriers and 9 low-risk women correctly reclassified; for unilateral cancer patients versus low-risk women, $\Delta\text{AUC} = 0.11$; 12 unilateral cancer women and 5 low-risk women correctly reclassified). However, the fusion classifiers performed comparable with the classifiers based on CNN features alone ($p = 0.3227$ for BRCA1/2 gene-mutation carriers versus low-risk women; $p = 0.3627$ for unilateral cancer patients versus low-risk women).

4 Discussion and Conclusion

In this FFDM study, we performed breast cancer risk assessment using features extracted with a conventional RTA method and with a deep CNN method that employed transfer learning. The results showed that the classification performances using features extracted from deep CNNs with transfer learning and conventional RTA method were comparable. By combining the outputs from the CNN-based method and the RTA-based method, a statistically significant improved performance was achieved as compared to the RTA method in terms of distinguishing between high- and low-risk groups for breast cancer risk assessment.

The fusion classifier, which combined the outputs from the CNN- and the conventional RTA-based method, yielded improved classification performance in the breast cancer risk assessment. It may be due to the fact that the features extracted from deep CNNs with transfer learning provided additional information regarding the mammographic parenchymal patterns. This observation was supported by the fair or slight kappa coefficients obtained between the outputs from two separate classifiers (which had merged either the CNN- or RTA-based features).

For the features extracted using deep CNNs with transfer learning, only the features from the first fully connected layer were investigated in this study, due to its relatively low dimensionality. Since this was a pretrained CNN, which had been trained with nonmedical images, its earlier layers would provide more generalizable features, and the features extracted from later layers were more specific to the original classification task. The outputs from earlier layers may be better with a larger dataset due to their even higher dimensionalities. In the future, both the low-level information extracted from early convolutional layers and the high-level information extracted from later layers will be investigated. In addition, a "fine-tuning" technique will be also explored, since other studies showed that using pretrained CNNs with fine-tuning can achieve improved performance.[27,44]

The conventional RTA features are more intuitive and can be relatively easy to relate to the characteristics of mammograms. However, features extracted from deep CNNs are not intuitive and difficult to interpret in terms of their clinical relevance. This needs to be further investigated.

There are several limitations in this study. The dataset size was relatively small (456 cases), which limited us to use a pretrained CNN as merely a feature extractor. A larger dataset is needed to perform fine-tuning. Also, the small dataset only allows us to perform a leave-one-case-out cross-validation evaluation. Given a larger dataset, splitting the dataset into training, validation, and testing would be more ideal for parameter optimization, model building, and robust performance evaluation.

In this preliminary study, we demonstrated that using features extracted with pretrained CNNs can achieve comparable performance to that using features extracted from a conventional RTA method in breast cancer risk assessment. The features extracted using CNNs may contain additional information in characterizing mammographic parenchymal patterns to the conventional RTA features. Deep learning has potential to help clinicians in assessing mammographic parenchymal patterns for breast cancer risk assessment.

Disclosures

M.L.G. is a stockholder in R2 Technology/Hologic and receives royalties from Hologic, GE Medical Systems, MEDIAN Technologies, Riverain Medical, Mitsubishi, and Toshiba. She is a cofounder of and stockholder in Quantitative Insights. H.L. receives royalties from Hologic. It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest that would reasonably appear to be directly and significantly affected by the research activities.

Acknowledgments

This work was partially funded by National Institutes of Health (NIH) CA195564, NIH CA189240, and the University of Chicago Metcalf program.

References

1. E. J. Feuer et al., "The lifetime risk of developing breast cancer," *JNCI J. Natl. Cancer Inst.* **85**(11), 892–897 (1993).
2. R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA Cancer J. Clin.* **66**(1), 7–30 (2016).
3. C. H. Lee et al., "Breast cancer screening with imaging: recommendations from the society of breast imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer," *J. Am. Coll. Radiol.* **7**(1), 18–27 (2010).
4. K. McPherson, C. M. Steel, and J. M. Dixon, "ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics," *BMJ* **321**(7261), 624–628 (2000).
5. J. N. Wolfe, "Breast patterns as an index of risk for developing breast cancer," *Am. J. Roentgenol.* **126**(6), 1130–1137 (1976).
6. J. N. Wolfe, "Risk for breast cancer development determined by mammographic parenchymal pattern," *Cancer* **37**(5), 2486–2492 (1976).
7. J. N. Wolfe, A. F. Saftlas, and M. Salane, "Mammographic parenchymal patterns and quantitative evaluation of mammographic densities: a case-control study," *Am. J. Roentgenol.* **148**(6), 1087–1092 (1987).
8. N. F. Boyd et al., "Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study," *JNCI J. Natl. Cancer Inst.* **87**(9), 670–675 (1995).
9. J. W. Byng et al., "Automated analysis of mammographic densities and breast carcinoma risk," *Cancer* **80**(1), 66–74 (1997).
10. Z. Huo et al., "Computerized analysis of digitized mammograms of BRCA1 and BRCA2 gene mutation carriers," *Radiology* **225**(2), 519–526 (2002).
11. N. F. Boyd et al., "Mammographic density and the risk and detection of breast cancer," *N. Engl. J. Med.* **356**(3), 227–236 (2007).
12. A. Manduca et al., "Texture features from mammographic images and risk of breast cancer," *Cancer Epidemiol. Biomarkers Prev.* **18**(3), 837–845 (2009).
13. J. Wei et al., "Association of computerized mammographic parenchymal pattern measure with breast cancer risk: a pilot case-control study," *Radiology* **260**(1), 42–49 (2011).
14. D. Kontos et al., "Analysis of parenchymal texture with digital breast tomosynthesis: comparison with digital mammography and implications for cancer risk assessment," *Radiology* **261**(1), 80–91 (2011).
15. H. Li et al., "Computerized analysis of mammographic parenchymal patterns on a large clinical dataset of full-field digital mammograms: robustness study with two high-risk datasets," *J. Digital Imaging* **25**(5), 591–598 (2012).
16. G. L. Gierach et al., "Relationships between computer-extracted mammographic texture pattern features and BRCA1/2 mutation status: a cross-sectional study," *Breast Cancer Res.* **16**(4), 424 (2014).
17. H. Li et al., "Comparative analysis of image-based phenotypes of mammographic density and parenchymal patterns in distinguishing between BRCA1/2 cases, unilateral cancer cases, and controls," *J. Med. Imaging* **1**(3), 031009 (2014).
18. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**(7553), 436–444 (2015).
19. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105 (2012).
20. W. Zhang et al., "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network," *Med. Phys.* **21**(4), 517–524 (1994).
21. B. Sahiner et al., "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images," *IEEE Trans. Med. Imaging* **15**(5), 598–610 (1996).
22. H. R. Roth et al., "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE Trans. Med. Imaging* **35**(5), 1170–1181 (2016).
23. F. Ciompi et al., "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box," *Med. Image Anal.* **26**(1), 195–202 (2015).
24. Y. Bar et al., "Deep learning with non-medical training used for chest pathology identification," *Proc. SPIE* **9414**, 94140V (2015).
25. B. van Ginneken et al., "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *12th Int. Symp. on Biomedical Imaging (ISBI)*, pp. 286–289, IEEE (2015).
26. R. K. Samala et al., "Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography," *Med. Phys.* **43**(12), 6654–6666 (2016).
27. H.-C. Shin et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imaging* **35**(5), 1285–1298 (2016).
28. B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *J. Med. Imaging* **3**(3), 034501 (2016).
29. H. Li et al., "Computerized analysis of mammographic parenchymal patterns for assessing breast cancer risk: effect of ROI size and location," *Med. Phys.* **31**(3), 549–555 (2004).
30. M. H. Gail, "Twenty-five years of breast cancer risk models and their applications," *JNCI J. Natl. Cancer Inst.* **107**(5), djv042 (2015).
31. M. Amadasun and R. King, "Textural features corresponding to textural properties," *IEEE Trans. Syst. Man Cybern.* **19**(5), 1264–1274 (1989).
32. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.* **SMC-3**(6), 610–621 (1973).
33. H. Li et al., "Power spectral analysis of mammographic parenchymal patterns for breast cancer risk assessment," *J. Digital Imaging* **21**(2), 145–152 (2008).
34. H. Li et al., "Fractal analysis of mammographic parenchymal patterns in breast cancer risk assessment," *Acad. Radiol.* **14**(5), 513–521 (2007).
35. W. Chen et al., "Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images," *Magn. Reson. Med.* **58**(3), 562–571 (2007).
36. Z. Huo et al., "Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: feature selection," *Med. Phys.* **27**(1), 4–12 (2000).
37. H. Li et al., "Computerized texture analysis of mammographic parenchymal patterns of digitized mammograms," *Acad. Radiol.* **12**(7), 863–873 (2005).
38. N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed., Wiley, New York (1998).
39. J. Hua et al., "Optimal number of features as a function of sample size for various classification rules," *Bioinformatics* **21**(8), 1509–1515 (2005).
40. C. E. Metz, "ROC methodology in radiologic imaging," *Invest. Radiol.* **21**(9), 720–733 (1986).
41. Metz ROC Software, http://metz-roc.uchicago.edu/MetzROC/software.
42. S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.* **6**, 65–70 (1979).
43. J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.* **20**(1), 37–46 (1960).
44. N. Tajbakhsh et al., "Convolutional neural networks for medical image analysis: full training or fine tuning?" *IEEE Trans. Med. Imaging* **35**(5), 1299–1312 (2016).

Hui Li has been working on quantitative imaging analysis on medical images for over a decade. His research interests include breast cancer risk assessment, diagnosis, prognosis, response to therapy, understanding the relationship between radiomics and genomics, and their future roles in precision medicine with both conventional and deep learning approaches.

Maryellen L. Giger is the A.N. Pritzker professor of radiology/medical physics at the University of Chicago, and for over 30 years, she has conducted research on computer-aided diagnosis and quantitative image analysis in the areas of breast cancer, lung cancer, prostate cancer, and bone diseases. She is the vicechair of radiology for basic science research at the University of Chicago.

Benjamin Q. Huynh currently works on applying deep learning methods to medical image analysis. His research interests include computational statistics, computer vision, and nonparametric Bayesian techniques with applications to biomedical tasks.

Natalia O. Antropova is a PhD student in the medical physics program at the University of Chicago. Her thesis work is focused on developing radiomics for breast cancer diagnosis and prognosis. In particular, her interests include applying deep learning methods to medical images for clinical decision-making.