Published OnlineFirst October 11, 2018; DOI: 10.1158/1078-0432.CCR-18-1115

Clinical Cancer Research

Deep Learning to Distinguish Recalled but Benign Mammography Images in Breast Cancer Screening

Sarah S. Aboutalib[1], Aly A. Mohamed[2], Wendie A. Berg[2,3], Margarita L. Zuley[2,3], Jules H. Sumkin[2,3], and Shandong Wu[4]

Abstract

Purpose: False positives in digital mammography screening lead to high recall rates, resulting in unnecessary medical procedures to patients and health care costs. This study aimed to investigate the revolutionary deep learning methods to distinguish recalled but benign mammography images from negative exams and those with malignancy.

Experimental Design: Deep learning convolutional neural network (CNN) models were constructed to classify mammography images into malignant (breast cancer), negative (breast cancer free), and recalled-benign categories. A total of 14,860 images of 3,715 patients from two independent mammography datasets: Full-Field Digital Mammography Dataset (FFDM) and a digitized film dataset, Digital Dataset of Screening Mammography (DDSM), were used in various settings for training and testing the CNN models. The ROC curve was generated and the AUC was calculated as a metric of the classification accuracy.

Results: Training and testing using only the FFDM dataset resulted in AUC ranging from 0.70 to 0.81. When the DDSM dataset was used, AUC ranged from 0.77 to 0.96. When datasets were combined for training and testing, AUC ranged from 0.76 to 0.91. When pretrained on a large nonmedical dataset and DDSM, the models showed consistent improvements in AUC ranging from 0.02 to 0.05 (all $P > 0.05$), compared with pretraining only on the nonmedical dataset.

Conclusions: This study demonstrates that automatic deep learning CNN methods can identify nuanced mammographic imaging features to distinguish recalled-benign images from malignant and negative cases, which may lead to a computerized clinical toolkit to help reduce false recalls. *Clin Cancer Res; 1–8. ©2018 AACR.*

Introduction

Mammography is clinically used as the standard breast cancer screening exam for the general population and has been shown effective in early detection of breast cancer and in reduction of mortality (1–3). High recall (asking a woman back for additional workup after a screening mammogram) rates are, however, a concern in breast cancer screening. On average, approximately 11.6% of women in the U.S. (over 5 million women annually) who are screened using digital mammography are recalled and over 70% of more than 1 million breast biopsies performed annually are benign (4, 5), resulting in unnecessary psychological stress, medical costs, and clinical workload. Thus, reducing false recalls from screening mammography is of great clinical significance.

Observer performance in breast cancer detection in screening mammography varies widely and is influenced by experience, volume, and subspecialization among other factors (6). One approach to improving mammographic interpretation is to provide the radiologist with powerful computerized tools to aid in image interpretation and decision-making. To help radiologists make more accurate decisions on whether to recall a patient, we focused on building computer-aided models/classifiers that can distinguish subtle imaging characteristics of mammograms from patients who were recalled but biopsied with benign results from mammograms read as negative from the onset, and mammograms from recalled cases with biopsy-proven breast cancer. Computerized classifiers can assist radiologists in predicting which patients/images may be recalled but most likely benign.

Deep learning (7) coupled with a big dataset has shown promising performance in many artificial intelligence applications and is entering the field of biomedical imaging (8–10). The main architecture of deep learning for image data is the convolutional neural network (CNN; ref. 11). The most distinguishing strength of the CNN is that it can automatically learn and hierarchically organize features from a large dataset without manual feature engineering (7, 12, 13) and outperformed, in many scenarios, traditional manual-engineered imaging features/descriptors. Although deep learning CNN can be used as an offline feature-extractor where features are coupled with traditional classifiers (14), it is more common to build end-to-end deep learning classification models (15, 16). Studies applying deep learning to digital mammography images have focused on mass and microcalcification detection and classification, distinction between tumor and benign/normal tissue (17–21), breast tissue segmentation (22), and classification of breast anatomy (9).

Recently, a number of groups competed under the Digital Mammography DREAM challenge (23) to distinguish malignant

[1]Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania. [2]Department of Radiology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania. [3]Magee-Womens Hospital of University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania. [4]Departments of Radiology, of Biomedical Informatics, of Bioengineering, and of Intelligent Systems, University of Pittsburgh, Pittsburgh, Pennsylvania.

Corresponding Author: Shandong Wu, University of Pittsburgh, 3362 Fifth Avenue, Pittsburgh, PA 15213. Phone: 412-641-2567; Fax: 412-641-2582; E-mail: wus3@upmc.edu

doi: 10.1158/1078-0432.CCR-18-1115

Published OnlineFirst October 11, 2018; DOI: 10.1158/1078-0432.CCR-18-1115

Translational Relevance

Breast cancer screening mammography is currently affected by high false recall rates resulting in unnecessary stress to patients, increased medical costs, and increased clinical workload. Deep learning convolutional neural networks (CNN) can be used to recognize the nuanced imaging features that distinguish recalled but benign mammography images, which may not be identifiable by human visual assessment. These imaging features can then be used for imaging interpretation and lead to computational tools that aid radiologists in distinguishing these images and thus help in the reduction of the false recall rate. In addition, with the ability of the CNN model to distinguish between negative and malignant images, deep learning can also perform well in computer-aided diagnosis of breast cancer.

from negative mammogram images. While important, distinguishing potentially recalled but biopsy benign images from both malignant and negative images represents a critical need and a practical approach to aid radiologists. The purpose of this study was to investigate end-to-end deep learning CNN models for automatic identification of nuanced imaging features to distinguish mammogram images belonging to negative, recalled-benign, and malignant cases aimed to improve clinical mammographic image interpretation and reduce unnecessary recalls.

Materials and Methods

We performed a retrospective study that received Institutional Review Board approval at our institution. This study was compliant to the Health Insurance Portability and Accountability Act and the U.S. Common Rule. Informed consent from patients was waived due to the retrospective nature.

Study cohort and datasets

We used two independent mammography datasets to develop and evaluate deep learning classifiers. These two independent mammogram datasets include a total of 3,715 patients and 14,860 images. A patient case typically contains a single patient exam with the standard four screening mammography views including left and right breast with craniocaudal (CC) and mediolateral oblique (MLO) views. Malignant images were taken from patients determined to have breast cancer based on the pathology results (only images of the cancer-affected breast were used). Negative images were from those who maintained a breast cancer–free status after at least a one-year follow-up. Recalled-benign images were taken from patients who were recalled based on the screening mammography exam but later determined as benign (biopsy-proven or diagnostically confirmed).

Full-field Digital Mammography dataset. This is a retrospective cohort of 1,303 patients (5,212 mammogram images) who underwent standard digital mammography screening (2007–2014) at our institution: 552 patients were evaluated as negative in the initial screen; 376 patients were recalled-benign; 375 patients were evaluated as positive for breast cancer (101 or 26.9% ductal carcinoma *in situ* and 274 or 73.1% invasive) based on pathology results.

Digital Database of Screening Mammography dataset. The Digital Database of Screening Mammography (DDSM) dataset (24–26) is a large collection of digitized film mammography images. A total of 9,648 images consisting of 2,412 patient cases were used from this dataset. Six-hundred and ninety-five cases were negative, 867 malignant, and 850 recalled-benign.

Deep learning approach for building classifiers

We built end-to-end two-class and three-class CNN models to investigate six classification scenarios: Five binary classifications: malignant versus recalled-benign + negative (i.e., recalled-benign and negative are merged into one class), malignant versus negative, malignant versus recalled-benign, negative versus recalled-benign, and recalled-benign versus malignant + negative (negative and malignant are merged into one class), as well as one triple classification: malignant versus negative versus recalled-benign.

The CNN used a modified version of the AlexNet (11, 27) model. The CNN structure consists of five convolutional layers (includes max-pooling in the 1st, 2nd, and 5th convolutional layers) followed by two fully connected layers and a fully connected output layer with a final softmax function. CNN model parameters were fixed in all experiments: batch size of 50 for stochastic gradient descent, a weight decay of 0.001, and a momentum of 0.9. For the learning rate, we started with 0.001 and dropped the rate by a factor of 10 every 2,500 iterations. To speed up training, rectified linear units were used as the activation function in place of traditional tangent sigmoid functions (7). To maximize performance and increase computational efficiency of the AlexNet network, images were preprocessed using standard techniques including histogram equalization, mean subtracting, and downsampling using standard bicubic interpolation to 227 × 227 pixels from original resolution.

Transfer learning was used to enhance model training. We pretrained all the CNN models with a large existing image dataset [ImageNet (28), 1.3 million nonmedical images] and then fine-tuned the models with our own mammography dataset. In a novel approach, we also tested the incremental transfer learning strategies: first pretraining the network using ImageNet and then continuing with the DDSM dataset, and finally fine-tuning and testing on the Full-field Digital Mammography (FFDM) dataset.

The CNN was trained with the goal of increasing variation of the data and avoiding overfitting. Internal validation is based on 6-fold cross validation of training data in the CNN model training phase, reducing overfitting and calibrating accuracy of the training process. External evaluation of the CNN models was performed using unseen testing data.

The deep learning network was implemented using the Caffe platform running on a desktop computer system with the following specifications: Intel Core i7-2670QM CPU@2.20GHZ with 8 GB RAM and a Titan X Pascal Graphics Processing Unit (GPU).

Evaluation and statistical analysis

An independent test cohort consisting of approximately 5% of our datasets was used for testing (95% data for training). We used this setting to maximize the amount of data for CNN model training. Table 1 summarizes the number of images used in our independent training and testing datasets in each experiment. As an additional robustness analysis, we also repeated the

Published OnlineFirst October 11, 2018; DOI: 10.1158/1078-0432.CCR-18-1115

Table 1. Number of training and testing images from the FFDM and digital DDSM datasets used for each experiment. The negative versus recalled-benign scenarios (negative vs. recalled-benign) have more data and thus are listed separately. For simplicity, all other scenarios (malignant vs. negative + recalled-benign, malignant vs. negative, malignant vs. recalled-benign, malignant vs. negative vs. recalled-benign, recalled-benign vs. malignant + negative) are listed under "Others." Both the total number of images in the scenario as well as the images per category are displayed.

		Number of training and testing images used			
		Training		Testing	
Dataset/experiment	Scenario	Total	Per category	Total	Per category
FFDM only (Train & Test)	Negative vs. recalled-benign	3,040	1,520	160	80
	Others	1,734	867	100	50
DDSM only (Train & Test)	Negative vs. recalled-benign	5,282	2,641	278	139
	Others	3,294	1,647	172	86
FFDM+DDSM (Train & Test)	Negative vs. recalled-benign	8,322	4,161	438	219
	Others	5,028	2,514	272	136
Pre-train on ImageNet and DDSM, test on FFDM	Negative vs. recalled-benign	3,040	1,520	160	80
	Others	1,734	867	100	50

experiments by using 10% and 15% of the data as independent sets for testing. In all settings, testing data did not include any images used in training.

We performed four different experiments/strategies utilizing the two datasets in testing the CNN model's performance in six classification scenarios. These four experiments included: (i) using the FFDM dataset only for training and testing; (ii) using the DDSM dataset only for training and testing, in order to determine how our CNN model performs on an independent dataset; (iii) using the FFDM and DDSM datasets together by combining the datasets (mixing up) for both training and testing, and finally (iv) using incremental transfer learning by pretraining the CNN model on the DDSM dataset after pretraining on the ImageNet dataset. Note that in all four experiments the base CNN models are pretrained on the ImageNet and fine-tuned with the FFDM dataset.

To explore clinical caution for reducing the risk of false negatives in recalling a woman, we performed additional classification experiments using CNN models to distinguish false negatives from recalled-benign cases and also from negative cases. Here, we used interval cancers as the false negative cases in the experiments.

The ROC curve (29) was generated and AUC was calculated as a metric of classification accuracy. For the triple classification, since ROC is a binary-class evaluation method, we followed the common practice in literature of generating an ROC curve for each binary-class combination and then reporting the average of the AUCs. Ninety-five percent confidence intervals (CI) were calculated for AUC values using bootstrapping methods (30). DeLong test (31) was used to evaluate the statistical significance when comparing differences of AUCs. All statistical tests were two-sided.

Results

Results on FFDM dataset

As shown in Fig. 1, we found that all categories can be well-distinguished (AUC ranging from 0.66 to 0.81) in the FFDM dataset. The identified imaging features between recalled-benign and negative were most distinguishing (AUC = 0.81; 95% CI, 0.75–0.88). Recalled-benign versus malignant + negative (AUC = 0.76; 95% CI, 0.67–0.86) had the second best performance, followed by malignant versus negative (AUC = 0.75; 95% CI 0.65–0.84). Malignant versus recalled-benign had performance

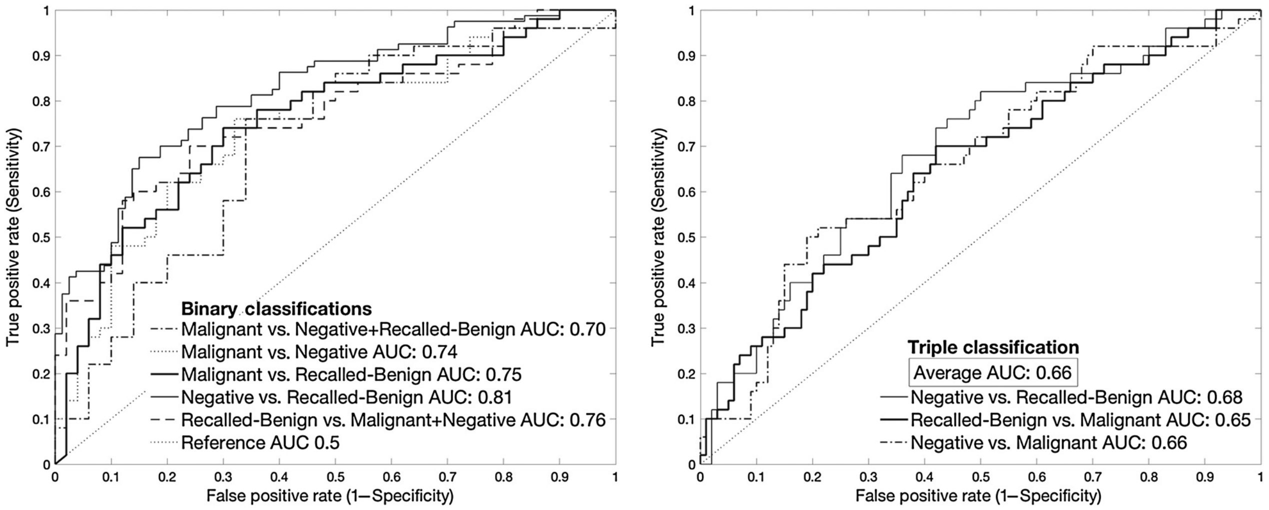

Figure 1.

Performance results for deep learning CNN models for classification on the FFDM dataset. Left, ROC curves for the binary classification scenarios and corresponding AUCs. Right, ROC curves for the triple-class classification scenario and averaged AUC.

Published OnlineFirst October 11, 2018; DOI: 10.1158/1078-0432.CCR-18-1115

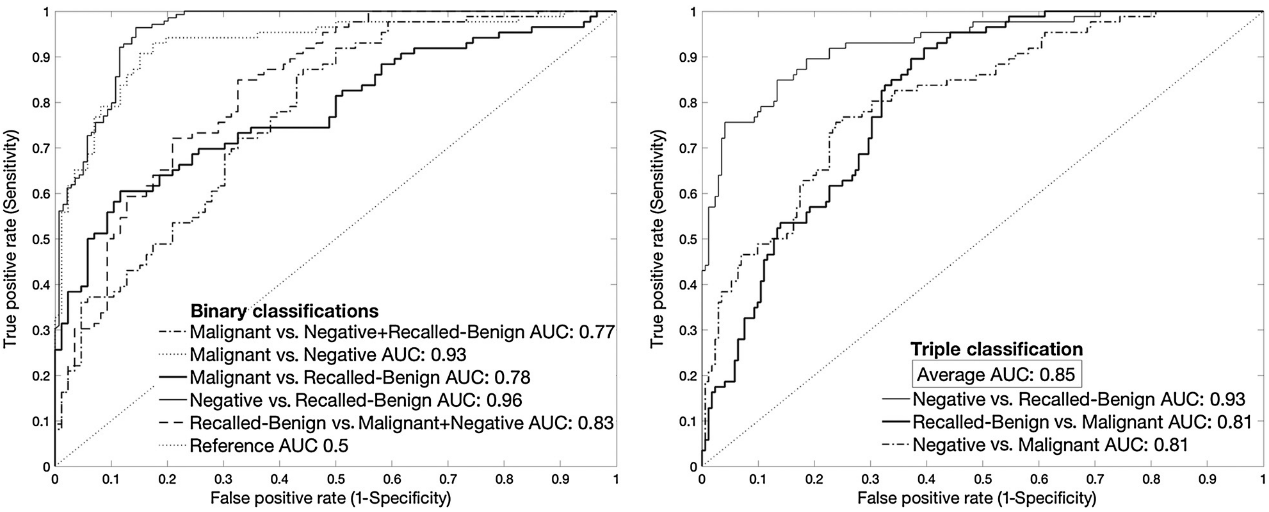

Figure 2.

Performance results for deep learning CNN models for classification on the DDSM dataset. Left, ROC curves for the binary classification scenarios and corresponding AUCs. Right, ROC curves for the triple-class classification scenario and averaged AUC.

result of AUC = 0.74 (95% CI, 0.65–0.84). Malignant versus negative + recalled-benign had AUC = 0.70 (95% CI, 0.60–0.80). Triple classification (malignant vs. negative vs. recalled-benign) had an average AUC of 0.66.

Results on DDSM dataset

Figure 2 shows the results on the DDSM dataset. We found AUC ranged from 0.77 to 0.90. Negative versus recalled-benign and malignant versus negative showed the best performances with AUC of 0.96 (95% CI, 0.94–0.98) and 0.93 (95% CI, 0.89–0.97), respectively. Triple classification had an average AUC of 0.85. Recalled-benign versus malignant + negative showed the next best performance at 0.83 (95% CI, 0.77–0.89). Malignant versus recalled-benign and malignant versus negative + recalled-benign showed similar performance with AUCs of 0.78 (95% CI, 0.71–0.85) and 0.77 (95% CI, 0.70–0.84), respectively.

Results on combined FFDM and DDSM datasets

Figure 3 shows performance results when the two datasets were combined (mixed together) for both training and testing. AUC ranged from 0.76 to 0.91. Best performance was observed for negative versus recalled-benign (AUC = 0.91; 95% CI, 0.89–0.94). With an AUC of 0.84 (95% CI, 0.79–0.88), malignant versus negative had second best performance, followed by malignant versus negative versus recalled-benign with an average AUC of 0.79, then recalled-benign versus malignant + negative

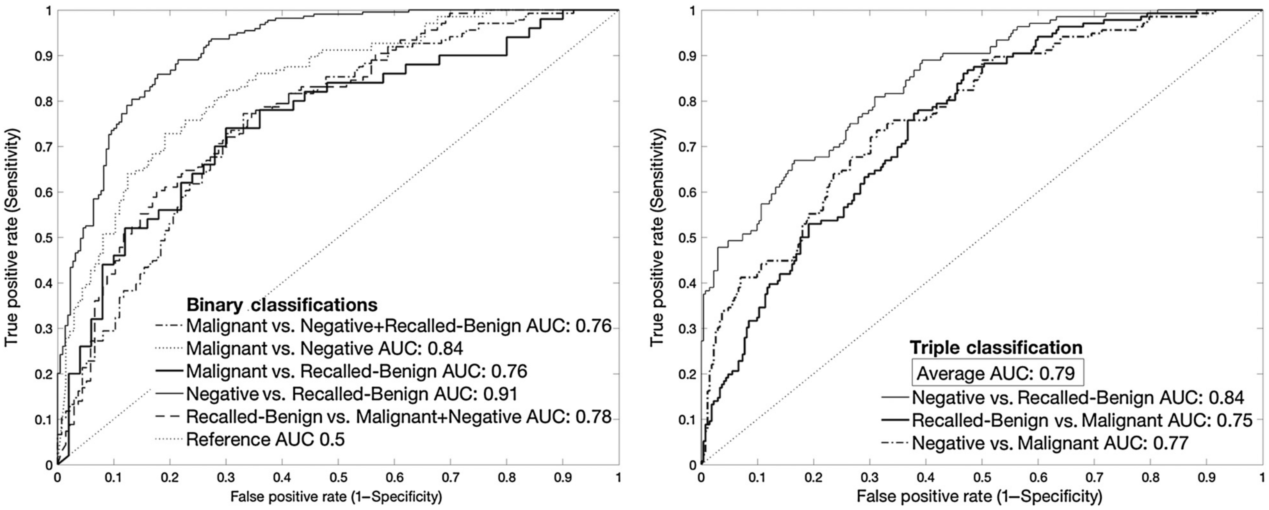

Figure 3.

Performance results for deep learning CNN models for classification using combined FFDM and DDSM datasets for training and testing. (Left) ROC curves for the binary classification scenarios and corresponding AUCs. Right, ROC curves for the triple-class classification scenario and averaged AUC.

Published OnlineFirst October 11, 2018; DOI: 10.1158/1078-0432.CCR-18-1115

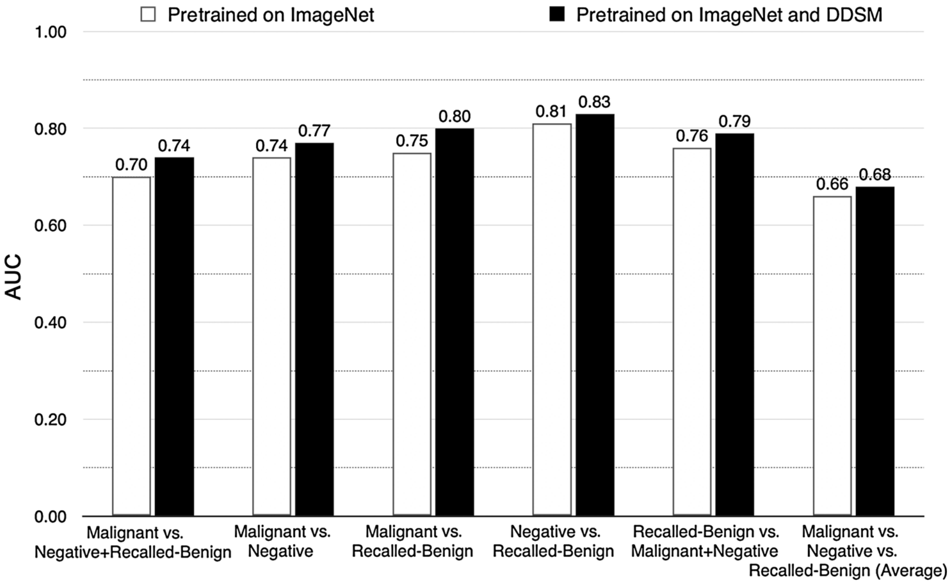

Figure 4.
Comparison of performance results of deep learning CNN models on different pretraining strategies: using original ImageNet pretrained model versus using model pretrained on ImageNet and DDSM dataset. All the AUCs were results based on training (fine-tuning) and testing on the FFDM dataset.

(AUC = 0.78; 95% CI, 0.73–0.83). Both malignant versus recalled-benign and malignant versus negative + recalled-benign had an AUC of 0.76 (95% CI, 0.70–0.82). It is observed that when FFDM and DDSM were combined for training and testing, overall classification showed AUC values between FFDM-only and DDSM-only results.

Incremental transfer learning using the DDSM dataset

DDSM dataset was used to further pretrain the base CNN network (after originally being pretrained on the ImageNet dataset), which was then fine-tuned with the FFDM dataset. Testing data was also from the FFDM dataset. As can be seen in Fig. 4, when compared with results using ImageNet pretrained model, all scenarios showed an increase in performance using this incremental transfer learning strategy, although not statistically significant ($P > 0.05$). Malignant versus recalled-benign showed the greatest increase in performance increasing from 0.75 to 0.80 (95% CI, 0.71–0.88, $P = 0.10$) or 5%, followed by malignant versus negative + recalled-benign (AUC = 0.70–0.74; 95% CI, 0.65–0.84; $P = 0.15$). Malignant versus negative and recalled-benign versus malignant + negative both had their AUC increase by 0.03, 0.74 to 0.77 (95% CI, 0.68–0.87; $P = 0.30$) and 0.76 to 0.79 (95% CI, 0.70–0.87; $P = 0.31$) respectively. Negative versus recalled-benign and the triple classification (malignant vs. negative vs. recalled-benign) had the least increase (2% or 0.02 in AUC, 0.66–0.68 for triple classification), although negative versus recalled-benign had the best performance of all the scenarios [AUC = 0.83 (95% CI 0.76–0.87; $P = 0.34$) compared with 0.81 for ImageNet pretrained model].

Robustness analysis

In the robustness analysis using 10% and 15% data for testing (Fig. 5), as expected, we see a slight decrease in the AUC performance in almost all the scenarios compared with those using 5% for testing. This is partly due to the reduced amount of data for training. We also found that the performances overall remained relatively stable, demonstrating the robustness of our CNN models on a varied amount of data for training and testing.

False negative analysis

The CNN models were further tested on correctly identifying false negative cases/images. In our FFDM dataset, there are 34 interval cancer patients examined as false negative cases. We retrained the CNN models (both the malignant vs. negative scenario and the malignant vs. recalled-benign scenario) on the FFDM dataset only and excluding those interval cancer patients, and then tested these models by inputting the false negative images to the models. Our results showed that 71.3% (by the malignant vs. negative model) and 63.6% (by the malignant vs. recalled-benign model) of the entire false negative images can be correctly identified by our models. Furthermore, when we retrained the CNN model by including 50% of the false negative cases and used the rest of the unseen 50% for testing, we see improved results, as expected, that is, 72.8% (by the malignant vs. negative model) and 68.4% (by the malignant vs. recalled-benign model) of the entire false negative images can be correctly identified. The CNN models trained by other strategies (i.e., experiments iii and iv in Section "Evaluation and statistical analysis") also achieved similar performance (results not shown).

Discussion

In this study, we present a novel investigation showing that automatic deep learning CNN methods can identify nuanced mammographic imaging features to distinguish negative, recalled-benign, and malignant images. We demonstrated that it is feasible to discover subtle imaging distinction for six classification scenarios on two different imaging datasets.

Among the six scenarios, negative versus recalled-benign showed the best performance. This scenario also had the greatest amount of data, which may have contributed to this improved performance. The distinction in this scenario implies that certain imaging features may result in recalled-benign images to be recalled rather than being determined negative in the first place. In general, the relatively higher AUCs in negative versus recalled-benign and also malignant versus recalled-benign indicate that there are imaging features unique to recalled-benign images that the CNN-based deep learning can identify and potentially use to

Published OnlineFirst October 11, 2018; DOI: 10.1158/1078-0432.CCR-18-1115

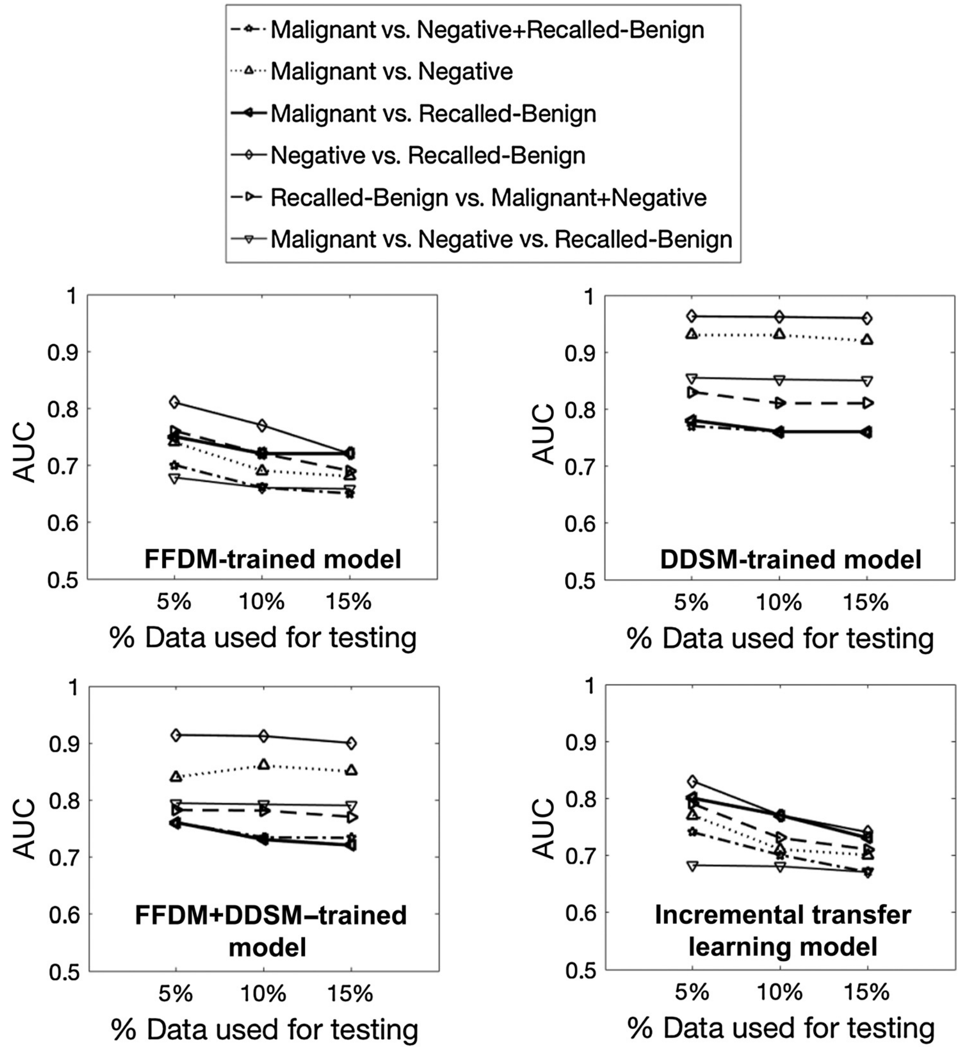

Figure 5.
Comparison of performance results using varying amounts (5%, 10%, and 15%) of testing data across all models: FFDM-trained model (top left); DDSM-trained model (top right); FFDM + DDSM–trained model (bottom left); incrementally pretrained CNN models in all scenarios (bottom right).

help radiologists in making better decisions on whether a patient should be recalled or is more likely a false recall. The six classification scenarios were designed to reveal different aspects of performance of the CNN models. It however remains to be determined which scenario, especially in terms of binary-classification or triple-classification, would be the most useful and meaningful choice in implementing a real-world CNN model to help radiologists make recall decisions in a clinical setting. There are noticeable variations in the classification performance between binary and triple classification. This is something to be further investigated in conjunction perhaps with reader studies to evaluate their clinical effects in depth. Of note, with the ability of the CNN model to distinguish between negative and malignant images, the results indicate that deep learning can also perform well in computer-aided diagnosis of breast cancer.

In terms of diagnosis (i.e., malignant vs. negative), literature has reported a great reader variability of radiologists in sensitivity and specificity in screening mammography (4). Using previously reported radiologists' overall sensitivity 86.9% as a reference threshold (4), our best deep learning model for classifying malignant versus negative yields a specificity of 87%, comparable with radiologists' overall specificity of 88.9%. This is encouraging for the current scale of our datasets. With enlarged datasets, we would expect further improvement in our model's performance. In terms of our major motivation of this work for reducing recalls, the lack of exact radiologists' performance data in the literature prevents us from making a reasonable comparison to our CNN models. We aim to further investigate this with a reader study in future work.

Our CNN models demonstrated encouraging results in the false negative analysis. Prevention of false negatives may be an important reason for the high recall rate in current clinical practice. By examining classifications of the interval cancer cases, it indicates a potential of the CNN models to help correctly identify as malignant a majority of false negative cases from recalled-benign cases or from negative cases. This also implies that some cases may be recalled due to certain imaging indications besides the intention of preventing false negatives. Further study into this issue on a larger cohort of false negative cases is of great clinical importance.

In terms of dataset, we started with a digital mammography dataset, FFDM, as it is the current standard screening

Published OnlineFirst October 11, 2018; DOI: 10.1158/1078-0432.CCR-18-1115

mammography examination. To demonstrate that our CNN models can be used on a dataset from an independent institution, we tested our models on the DDSM dataset. The DDSM alone showed the best performance overall. This may be due to the larger dataset size or something intrinsic to the characteristics of the DDSM dataset. When the FFDM and DDSM datasets were combined, the overall improved performance was observed in comparison with using FFDM alone. This indicates that our CNN model is robust to an external dataset and its performance can be further improved by including additional data.

Because the DDSM dataset is based on digitized film mammography images and current clinical practice has moved to digital mammography, we wanted to determine the best use of the DDSM to improve results on the FFDM dataset. The most straightforward way was simply to combine it with the FFDM datasets, which improved results when testing on the combined datasets. More importantly, we demonstrated a novel approach of incremental transfer learning using the DDSM dataset, which enhanced the performance consistently on the FFDM dataset in all six scenarios. The incremental transfer learning we used was a two-phase transfer learning approach using two different datasets consecutively (i.e., ImageNet then DDSM). The improvement may be due to using DDSM to fine-tune the model's weights rather than directly in training, allowing the noise from being a different type of mammography dataset to be dampened. It could also be related to the strengths of DDSM as a medical imaging modality that is close to the target modality (i.e., digital mammography) in fine-tuning the weights learned from pretraining on the nonmedical imaging dataset of ImageNet. Many studies have shown the benefit of transfer learning in medical domains with limited data (32). Our results in this study provide deeper insights in developing more optimized transfer learning strategies. However, the incremental transfer learning and the observations made here needs to be evaluated by further analyses and comparative studies in future work.

In comparing the various deep/transfer learning strategies, we show that adding/mixing a larger independent dataset, even if it is not exactly the same imaging modality, helped improve the CNN models in our classification tasks. Furthermore, using this kind of additional dataset in an incremental transfer learning approach has shown a trend of boosted model performance. Although the AUC increases are not statistically significant, additional data will help further evaluate this finding. As to mixing two datasets or using transfer learning, it depends on the specific scenarios and their actual classification performance. Although we have shown encouraging results in this study, we believe this is still an open question meriting further investigation.

In traditional computer-aided detection or diagnosis, the models are usually based on predefined features, which require preemptive determination of which features will work best for the task at hand. In contrast, with the deep learning method we utilized, predefinition of the imaging features is not necessary and are learned automatically from labeled data. Deep learning allows nuanced features to be determined by the learning algorithm for the targeted task, where intrinsic features that may not be identifiable by human visual assessment can be automatically identified and used for imaging interpretation. This study illustrated the encouraging effects of such automatic feature identification by deep learning.

Our study has some limitations. Although we have two independent datasets, additional datasets, especially another digital mammography dataset, can further bolster the evaluation of the deep learning models. We are currently trying to obtain external datasets for such experiments. Also, it is still not clear why the DDSM performed substantially better than the FFDM dataset; this requires further exploration. Of note, digital breast tomosynthesis is increasingly being used in clinical practice, and it has been shown to reduce recall rates (33); thus, incorporation of tomosynthesis data in our study will be important future work. Although we utilized AlextNet CNN structures in this study, comparison with other network structures such as Residual Network (34), VGG (35), or GoogLeNet (36) will be useful to gain further insights on the potential of deep learning.

Finally, it would be useful to present to radiologists the subtle imaging features found by our CNN models in distinguishing the different groups. However, we are not yet able to clearly visualize and clinically interpret what the identified nuanced imaging features are for recalled but benign images or for the other classification categories. At this phase, deep learning is often referred to as a "black-box" due to the lack of interpretability of the identified features. A current area of investigation we are exploring is to visualize the CNN-identified features to be more intuitively perceived by radiologists. The complexity of deep learning network structures, parameters, and data evolving process across different network layers, however, make feature visualization very complicated, requiring in-depth research. Further technical advancement in this active research area is expected to contribute to addressing this important issue.

In summary, we showed that the three different imaging reading categories (malignant, negative, and recalled-benign) could be distinguished using our deep learning–based CNN models. We believe our study holds great potential to incorporate deep learning–based artificial intelligence into clinical workflow of breast cancer screening to improve radiologist interpretation of mammograms, ultimately contributing to reducing false recalls.

Disclosure of Potential Conflicts of Interest

J.H. Sumkin reports receiving commercial research grants from GE and Hologic. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: J.H. Sumkin, S. Wu
Development of methodology: S.S. Aboutalib, A.A. Mohamed, S. Wu
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): S.S. Aboutalib, M.L. Zuley, S. Wu
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): S.S. Aboutalib, A.A. Mohamed, S. Wu
Writing, review, and/or revision of the manuscript: S.S. Aboutalib, A.A. Mohamed, W.A. Berg, M.L. Zuley, J.H. Sumkin, S. Wu
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): J.H. Sumkin, S. Wu
Study supervision: W.A. Berg, S. Wu

Acknowledgments

This work was supported by NIH/NCI grants (#1R01CA193603, #3R01CA193603-03S1, and #1R01CA218405; to S. Wu, A.A. Mohamed, W.A. Berg, and M.L. Zuley), a National Library of Medicine (T15LM007059) grant (to S.S. Aboutalib), a Radiological Society of North America (RSNA) Research Scholar Grant (#RSCH1530; to S. Wu), a University of Pittsburgh Cancer Institute Precision Medicine Pilot Award from The Pittsburgh Foundation (#MR2014-77613; to S. Wu, A.A. Mohamed), and a University of

Pittsburgh Physicians (UPP) Academic Foundation Award (to S. Wu). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received April 13, 2018; revised June 19, 2018; accepted July 31, 2018; published first October 11, 2018.

References

1. Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, et al. Efficacy of breast cancer screening by age: new results from the Swedish two-county trial. Cancer 1995;75:2507–17.
2. U.S. Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2016;164:279–96.
3. Coldman A, Phillips N, Wilson C, Decker K, Chiarelli AM, Brisson J, et al. Pan-Canadian study of mammography screening and mortality from breast cancer. J Natl Cancer Inst 2014;106:pii:dju261.
4. Lehman D, Arao RF, Sprague BL, Lee JM, Buist DSM, Kerlikowske K, et al. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium constance. Radiology 2017;283:49–58.
5. Silverstein MJ, Lagios MD, Recht A, Allred DC, Harms SE, Holland R, et al. Image-detected breast cancer: state of the art diagnosis and treatment. J Am Coll Surg 2005;201:586–97.
6. Elmore JG, Wells CK, Howard DH. Does diagnostic accuracy in mammography depend on radiologists' experience? J Womens Health 1998;7:443–9.
7. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
8. Wang D, Khosla A, Gargeya R, Irshad H, Beck A. Deep learning for identifying metastatic breast cancer. Ithaca, NY: Cornell University; 2016.
9. Dubrovina A, Kisilev P, Ginsburg B, Hashoul S, Kimmel R. Computational mammography using deep neural networks. Comp Methods Biomech Biom Eng: Imaging Visualization 2016;6:1–5.
10. Huval B, Coates A, Ng A. Deep learning for class-generic object detection. Ithaca, NY: Cornell University; 2013.
11. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Informat Proc Syst 2012;25:1106–14.
12. Litjens G, Kooi T, Bejnordi BE. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
13. Deng L, Yu D. Deep learning: methods and applications. Foundat Trends Signal Process 2014;7:197–387.
14. Li H, Giger M, Huynh BQ, Antropova N. Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. J Med Imaging 2017;4:041304.
15. Carneiro G, Nascimento J, Bradely AP. Unregistered multiview mammogram analysis with pre-trained deep learning models. Int Conf Med Image Computing Computer-Assisted Intervention 2015;9351:652–60.
16. Mohamed A, Berg W, Peng H, Luo Y, Jankowitz R, Wu S. A deep learning method for classifying mammographic breast density categories. Med Phys 2017;45:314–21.
17. Dhungel N, Carneiro G, Bradley AP. Automated mass detection in mammograms using cascaded deep learning and random forests. In: Proceedings of the 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA); 2015 Nov 23–25; Adelaide, SA, Australia. 2015.
18. Ertosun MG, Rubin DL. Probabilistic visual search for masses within mammography images using deep learning. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2015 Nov 9–12; Washington, DC; 2015.
19. Levy D, Jain A. Breast mass classification from mammograms using deep convolutional neural networks. Ithaca, NY: Cornell University; 2016.
20. Arevalo J, Gonzalez FA, Ramos-Poll R, Oliveira JL, Lopez MAG. Representation learning for mammography mass lesion calssification with convolutional neural networks. Comp Methods Programs Biomed 2016;127:248–57.
21. Mordang JJ, Janssen T, Bria A, Kooi T, Gubern-Merida A, Karssemeijer N. Automatic microcalcification detection in multivendor mammography using convolutional neural networks. Int Workshop Digital Mammograph. 2016;9699:35–42.
22. Petersen K, Nielsen M, Diao P, Karssemeijer N, Lillholm M. Breast tissue segmentation and mammographic risk scoring using deep learning. Int Workshop Digital Mammography 2014;8539:88–94.
23. Dream Challenges. Digital Mammography DREAM Challenge. 2016–2017. Available from: https://www.synapse.org/Digital_Mammography_DREAM_Challenge.
24. Heath M, Bowyer K, Kopans D, Kegelmeyer P Jr, Moore R, Chang K, et al. Current status of the digital database for screening mammography. Proc 4th Int Workshop Digital Mammography 1998;13:457–60.
25. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer PW. The digital database for screening mammography. In Proceedings of the Fifth International Workshop on Digital Mammography; 2001. Madison, WI: Medical Physics Publishing; 212–8.
26. Sharma A. DDSM Utility. GitHub. 2015. Available from: https://github.com/trane293/DDSMUtility.
27. Yangqing J, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. In: https://github.com/BVLC/caffe/tree/master/models. arXiv preprint arXiv:1408.5093. 2014.
28. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 June 20–25; Miami FL; 2009.
29. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986;21:720–33.
30. Robin X, Turck N, Hainard A. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77.
31. DeLong E, DM D, Clarke-Pearson D. Comparing areas under two or more correlated reciever operating characteristics curves: a nonparamentric approach. Biometrics 1988;44:837–45.
32. Zhou Z, Shin J, Zhang L. Fine-tuning convolutional neural network for biomedical image analysis: actively and incrementally. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 July 21–26; Honolulu, HI; 2017.
33. McDonald E, Oustimov A, Weinstein S, Synnestvedt M, Schnall M, Conant E. Effectiveness of digital breast tomosynthesis compared with digital mammography: outcomes analysis from 3 years of breast cancer screening. JAMA Oncol 2016;2:737–43.
34. He K, Zhang X, Shaoqing R, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 June 27–30; Las Vegas, NV; 2016.
35. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Ithaca, NY: Cornell University; 2014.
36. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Ithaca, NY: Cornell University; 2014.

Published OnlineFirst October 11, 2018; DOI: 10.1158/1078-0432.CCR-18-1115

AACR American Association for Cancer Research

Clinical Cancer Research

Deep Learning to Distinguish Recalled but Benign Mammography Images in Breast Cancer Screening

Sarah S. Aboutalib, Aly A. Mohamed, Wendie A. Berg, et al.

Clin Cancer Res Published OnlineFirst October 11, 2018.