Do pre-trained deep learning models improve computer-aided classification of digital mammograms?

Aboutalib, Sarah, Mohamed, Aly, Zuley, Margarita, Berg, Wendie, Luo, Yahong, et al.

Sarah S. Aboutalib, Aly A. Mohamed, Margarita L. Zuley, Wendie A. Berg, Yahong Luo, Shandong Wu, "Do pre-trained deep learning models improve computer-aided classification of digital mammograms?," Proc. SPIE 10575, Medical Imaging 2018: Computer-Aided Diagnosis, 1057523 (27 February 2018); doi: 10.1117/12.2293777

SPIE.

Event: SPIE Medical Imaging, 2018, Houston, Texas, United States

Do pre-trained deep learning models improve computer-aided classification of digital mammograms?

Sarah S. Aboutalib[a], Aly A. Mohamed[b], Margarita L. Zuley[b], Wendie A. Berg[b], Yahong Luo[d], Shandong Wu[*a,b,c]

[a]Department of Biomedical Informatics, [b]Department of Radiology, [c]Department of Bioengineering, University of Pittsburgh,
3362 Fifth Ave, Pittsburgh, PA 15213, USA;

[d]Department of Radiology,
Liaoning Cancer Hospital & Institute,
44 Xiaoheyan Rd, Dadong District, Shenyang City, Liaoning Province, 110042, China

ABSTRACT

Digital mammography screening is an important exam for the early detection of breast cancer and reduction in mortality. False positives leading to high recall rates, however, results in unnecessary negative consequences to patients and health care systems. In order to better aid radiologists, computer-aided tools can be utilized to improve distinction between image classifications and thus potentially reduce false recalls. The emergence of deep learning has shown promising results in the area of biomedical imaging data analysis. This study aimed to investigate deep learning and transfer learning methods that can improve digital mammography classification performance. In particular, we evaluated the effect of pre-training deep learning models with other imaging datasets in order to boost classification performance on a digital mammography dataset. Two types of datasets were used for pre-training: (1) a digitized film mammography dataset, and (2) a very large non-medical imaging dataset. By using either of these datasets to pre-train the network initially, and then fine-tuning with the digital mammography dataset, we found an increase in overall classification performance in comparison to a model without pre-training, with the very large non-medical dataset performing the best in improving the classification accuracy.

Keywords: Breast cancer, screening mammography, recalls, deep learning, transfer learning

1. INTRODUCTION

For the early detection of breast cancer, digital mammography is clinically used as the standard breast cancer screening exam for the general population and has been shown effective in the reduction of mortality[1]. However, approximately 10% of women in the U.S. (over 5 million women annually) who are screened are recalled (asking a woman back for additional workup) with 80% of breast biopsies performed annually being benign[2,3]. This high recall is a difficulty to breast screening causing unnecessary psychological stress, clinical workload, and medical costs. Aiding radiologists in the interpretation of these digital mammograms during clinical readings is thus an important need, especially considering the variability in reading performances of individual radiologists in breast cancer detection in screening mammography due to experience, reader volume, and a reader's subspecialty, among other factors[5].

In order to aid radiologists in better distinguishing mammograms of patients with biopsy-proven breast cancer, those who are read negative at the onset, and those that are recalled but biopsy benign, computerized tools to identify imaging features unique to each of those scans can be used to enhance the radiologist's reading capability and decision making. Thus, our study focuses on developing computer-aided classifies to distinguish mammogram imaging characteristics. One aspect of building these classifiers is to increase the model's accuracy so that improve their clinical utility in helping reduce unnecessary recalls.

* wus3@upmc.edu; phone 1 412-641-2567;

Medical Imaging 2018: Computer-Aided Diagnosis, edited by Nicholas Petrick, Kensaku Mori, Proc. of SPIE Vol. 10575, 1057523 · © 2018 SPIE · CCC code: 1605-7422/18/$18 · doi: 10.1117/12.2293777

Deep learning is entering the field of biomedical imaging[6,7,8] after having shown promising results in many artificial intelligence applications when coupled with big data. Studies utilizing mammography images include breast anatomy classification[9], lesion diagnosis and discrimination[10-14], discrimination of masses and micro-calcifications and their combination[15], tissue/mass segmentation and risk scoring[16], Breast Imaging Reporting and Data System (BI-RADS) breast density category classification[17], and so on. Some methods utilize deep learning for mass detection or feature extraction in the recognition process[18,19] or for recognition using an underlying mass detector or region of interest[20,21].

The convolutional neural network (CNN)[22] is the main architecture of deep learning for image data. CNNs are biologically-inspired multi-layer neural networks that can automatically learn and hierarchically organize features from a large dataset without manual feature engineering. When limited data is available, deep learning models can be pre-trained with other larger datasets and then fine-tuned using the target dataset, a process called transfer learning. In this study, we investigated how pre-trained deep learning models may improve the accuracy of CNN classifiers. More specifically, we evaluated the effect of transfer learning to improve the classification of digital mammography scans in two scenarios, i.e., pre-training a CNN model by using: (1) a larger analog scanned but digitized film mammography dataset and (2) an extremely large but non-medical imaging dataset. Improvement in CNN model performance by utilizing transfer learning can potentially lead to better computer-aided tools for radiologists in reducing the false recalls.

2. METHODS

2.1 Overview

In order to test the effect of pre-trained deep learning models on the classification of digital mammography images, we utilized two datasets (see below) to pre-train the network and compared it with a deep learning network that was not pre-trained. Testing was done on a digital mammography dataset described below in four classification scenarios: Positive vs Negative+Recalled-Benign (Recalled-benign and Negative are treated as one class with equal representation within that class), Positive vs Negative, Positive vs Recalled-Benign, and Negative vs Recalled-Benign. Positive images were taken from patients that were determined to have breast cancer based on pathology results, only images of the cancer-affected breast were used. Negative images were taken from breast images whose patient was determined to be breast cancer free 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 benign based on pathology results.

2.2 Study cohort and datasets

We performed a retrospective study that was compliant to the Health Insurance Portability and Accountability Act (HIPAA) and received Institutional Review Board (IRB) approval at our institution. Informed consent from patients was waived due to the retrospective nature. Below we give a description of the datasets used, including the target dataset and two pre-training datasets:

Target dataset: Full-field Digital Mammography (FFDM) Dataset:

This study cohort includes 1303 patients (5234 mammogram images) who underwent general population digital mammography screening at our institution during 2007-2014. We used the post-processed (i.e., "FOR PRESENTATION") images. Both craniocaudal (CC) and mediolateral oblique (MLO) image views were used for all patients. For training, 1734 images were used for training for each of the classification scenarios listed above (867 images for each category), except for Negative vs Recalled-Benign (which had more data), in which 3040 images were used for training (1520 for each category). A balanced number of images were used for training in each category in order to maintain a balanced representation and reduce potential bias in the learning. This dataset was used as the target dataset for examining (including both training and testing) the CNN-based classification accuracy.

Pre-training dataset 1: Digital Database of Screening Mammography (DDSM) Dataset:

The DDSM dataset[23,24,25] is a publically available large collection of digitized film mammography images on the order of 10,000 images, consisting of 2620 patient cases, where a case typical contains a single patient exam with the standard four screening mammography views: Left CC, right CC, left MLO, and right MLO. A total of 9648 images consisting of 2412 patient cases were used from this dataset, including 695 cases labeled as normal (i.e. negative), 867 labeled as breast cancer (i.e. positive), and 850 cases as benign (i.e. recalled, but determined benign through pathology). This dataset was used as a pre-training dataset.

Pre-training dataset 2: ImageNet Dataset:

The ImageNet Dataset[26] is a non-medial image dataset used in the Large Scale Visual Recognition Challenge (ILSVRC). We utilized the 2010 dataset consisting of 1.3 million labeled images of objects and animals including 1000 different classes. This dataset was used as a pre-training dataset.

2.3 Equipment

The deep learning network was implemented using the Caffe platform[27] running on a system with the following specifications: Intel® Core™ i7-2670QM CPU@2.20GHZ with 8 GB RAM and a Titan X Pascal GPU.

2.4 Deep learning approach

A two-class CNN model was used to classify four scenarios of two-category classification outlined above (e.g., "Recalled-Benign vs Negative"). The CNN used a modified version of the AlexNet model[22,27] with no relighting data augmentation and the order of the pooling and normalization layers switched for better performance on the Caffe deep learning platform[27]. The CNN structure consists of five convolutional layers, three max-pooling layers, and three fully connected layers with a final 2-way softmax function. The two-class CNN model was constructed as an end-to-end system aiming at classifying each of the two-category scenarios listed above.

The CNN was trained with the goal of increasing the variation of the data and avoiding overfitting. Rectified linear units (ReLU) were used as the activation function in place of the traditional tangent function and the sigmoid function. Key steps in the training process include:

(1) Run a histogram equalization on all training images to calibrate the contrast.

(2) Resize all images to 227x227, the standard image size for deep learning on AlexNet for computational efficiency.

(3) Generate the mean image of the training data and subtract from each input image to ensure feature pixel has zero mean.

(4) Apply 6-fold cross-validation in the CNN model training phase to calibrate accuracy of the training process and prevent overfitting.

In the model training process, the optimization of the hyperparameters was performed using a stochastic gradient descent (SGD) method with batch size of 50. In our configuration (weight decay of 0.001 and a momentum of 0.9), we started with a learning rate of 0.001 and dropped the learning rate by a factor of ten every 2500 iterations. These parameters were fixed in all the experiments.

2.5 Transfer learning

To maximize performance, a well-known technique, transfer learning, was used that takes weights from a previously trained model usually using a larger existing dataset independent from the target dataset as the initial weights in constructing the target CNN models.

We started with a CNN model without any pre-training and used this as a base model for comparison. Two CNN models were then pre-trained with the DDSM and ImageNet datasets, respectively, and each was fine-tuned by the target dataset (i.e., the FFDM dataset). In order to pre-train a CNN model, the same network topology was trained using the pre-training datasets (i.e., the DDSM and ImageNet). The weights from the pre-trained models are then saved and used as the initial weights in constructing the two final CNN models, where these weights are further fine-tuned by continuing the backpropagation using the target dataset.

2.6 Evaluation

An independent cohort consisting of approximately 5% of the target datasets was used for testing, so no training images were used in the test set. Specifically, the test set includes 160 images (80 for each category) of the FFDM dataset for the Negative vs Recalled-Benign scenario, while 100 images (50 for each category) of the FFDM dataset are included for the rest of the four scenarios. Table 1 summarizes the numbers of training and testing images for each scenario.

The receiver operative characteristic (ROC) was generated and the area under the curve (AUC) was calculated as a metric of the classification accuracy. Each of the three CNN models (Not pre-trained, pre-trained with DDSM, and pre-trained with ImageNet) were tested in all four scenarios (Positive vs Negative+Recalled-Benign, Positive vs Negative, Positive vs Recalled-Benign, and Negative vs Recalled-Benign) and their AUC's were compared.

3. RESULTS

Figure 1 compared the classification results for the three CNN models. When not pre-trained, AUC of the base CNN model ranged from 0.64 to 0.67 with Negative vs Recalled-Benign having the largest value, followed by Positive vs Recalled-Benign (AUC 0.66), then Positive vs Negative (AUC 0.65) and lastly Positive vs Negative+Recalled-Benign (AUC 0.64).

The classification results when the CNN model was pre-trained with the DDSM dataset showed AUC values ranged from 0.69 to 0.72 with Positive vs Negative and Negative vs Recalled-Benign having the best performance (AUC 0.72), followed by Positive vs Recalled-Benign (AUC 0.71), and lastly Positive vs Negative+Recalled-Benign (AUC 0.69). When compared with the results of the CNN model without pre-training, we see an average improvement of 8.5%, with improvements ranging from 0.05 to 0.07 in absolute values or 7.7% to 10.8% percent increase.

The AUC of the CNN model pre-trained with the ImageNet dataset ranged from 0.73 to 0.82, with Negative vs Recalled-Benign having the best performance (AUC 0.82), followed by Positive vs Recalled-Benign (AUC 0.78), then Positive vs Negative (AUC 0.74) and lastly, Positive vs Negative+Recalled-Benign (AUC 0.73). There was an average improvement of 17.1% in performance over the CNN that was not pre-trained, ranging from 13.8% to 22.4% or 0.09 to 0.15 in absolute values. There was also an improvement as compared to the network pre-trained with the DDSM dataset, although not as large as the network that was not pre-trained. Average percent increase between the DDSM pre-trained model and the ImageNet pre-trained model was 7.7% with a range of 0.02 to 0.10 in absolute values, and 2.8% to 12.2% percent increase.

Table 1. Number of training and testing images from the full-field mammography dataset (FFDM) used as target dataset for fine-tuning and evaluation.

	Scenario	**Training**		**Testing**	
		Total	Per Category	Total	Per Category
Target Dataset: FFDM	Negative vs Recalled-Benign	3040	1520	160	80
	Positive vs Negative Positive vs Negative+Recalled-Benign Positive vs Recalled-Benign	1734	867	100	50

4. DISCUSSION

In this study we looked into the effect of pre-training using both medical and non-medical datasets to improve deep learning results for digital mammography image classification. We investigated pre-training using two different datasets: (1) a larger related but originally analog version of mammography data, i.e. digitized film mammograms and (2) a much larger non-medical imaging dataset. We showed that in both cases results in the classification of digital mammograms could be improved with pre-training using either dataset, but with the larger non-medical results getting better results. These findings will be instrumental in developing deep learning-based computer-aided tools for help radiologists making more accurate decision-making on whether to recall a patient.

It is very interesting to see a clear improvement when the DDSM dataset was utilized for pre-training, meaning that digitized film mammography can be used to improve performance on digital mammography scans. Since current clinical practice has moved to digital mammography scans, we can still utilize older digitized film mammography datasets to boost performance on more current digital mammography scans. The best performance, however, was found when a very large (>1 million) dataset was used, even though the images were non-medical. This may be due to the fact that the amount of data is a major driving force for good performance on a deep learning network. This finding will need further studies to examine.

For future work, we would like to explore how additional variation in pre-training strategies can be used to further boost performance. In addition, we will expand our datasets to include more mammography images including digital mammography images from other institutions for robustness analyses of our findings. We also plan to do a detailed comparison of performance when using other deep learning network architectures, parameters, and platforms.

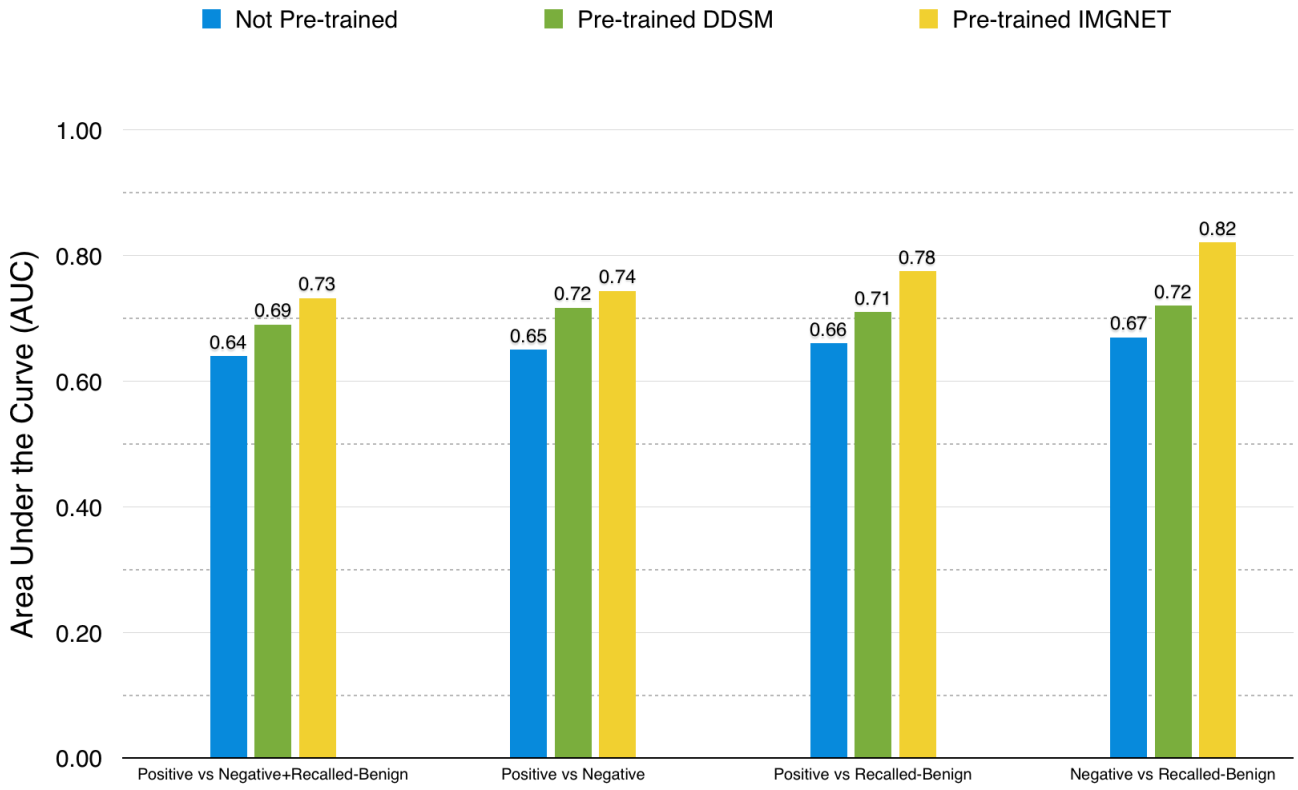

Figure 1. Performance results (area under the receiver operating characteristic) for deep learning CNN models without pre-training, pre-training with DDSM dataset, and pre-training with ImageNet dataset in the classification of digital mammography images in four classification scenarios.

5. CONCLUSION

We presented a novel investigation into the use of pre-trained deep learning networks for the classification of mammogram images focusing on four classification scenarios. Classifiers on these scenarios have the potential to contribute to generating computer-aided tools to help reduce false recalls. In addition, we tested the utilization of transfer learning to boost deep learning performance by pre-training the network with large existing datasets. In all cases, distinction between the categories (positive, negative, and recalled-benign) was encouraging. A significant improvement in performance was observed, however, in the networks where pre-training was done. Our study provides insight into how computer-aided systems to aid radiologists in the classification of digital mammography images can be improved by first pre-training the CNN model then fine-tuning the model using a target dataset.

ACKNOWLEDGEMENTS

This work was supported by a National Institutes of Health (NIH)/National Cancer Institute (NCI) R01 grant (#1R01CA193603), a R01 Supplement grant (#3R01CA193603-03S1), a National Library of Medicine (T15 LM007059) grant, a Radiological Society of North America (RSNA) Research Scholar Grant (#RSCH1530), a University of Pittsburgh Cancer Institute Precision Medicine Pilot Award from The Pittsburgh Foundation (#MR2014-77613), and a University of Pittsburgh Physicians (UPP) Academic Foundation Award. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

REFERENCES

[1] Tabar, L., Fagerberg, G., Chen, HH., et al., "Efficacy of breast cancer screening by age: new results from the Swedish two-county trial", Cancer 75, 2507-2517 (1995).

[2] Silverstein, MJ., Lagios, MD., Recht, A., et al., "Image-detected breast cancer: state of the art diagnosis and treatment", J Am Coll Surg. 201(4), 586-97 (2005).

[3] Weaver, D.L., Rosenberg, R.D.,et al.,"Pathologic findings from the Breast Cancer Surveillance Consortium: population-based outcomes in women undergoing biopsy after screening mammography",Cancer106, 732-42 (2006)

[4] Berg, W.A., D'Orsi, C.J., Jackson, V.P., et al., "Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography?", Radiology 224, 871–880 (2002).

[5] Elmore, J.G., Wells, C.K., Howard, D.H., "Does diagnostic accuracy in mammography depend on radiologists' experience?", J Womens Health 7, 443–449 (1998).

[6] Wang, D., Khosla, A., et al., "Deep Learning for Identifying Metastatic Breast Cancer", arXiv:1606.05718 (2016).

[7] Dubrovina, A., Kisilev, P., Ginsburg, B., Hashoul, S., and Kimmel, R., "Computational mammography using deep neural networks", Comp Methods in Biomechanics and Biomed Engineering: Imaging & Visualization, 1-5 (2016).

[8] Huval, B., Coates, A. & Ng, A., "Deep learning for class-generic object detection", arXiv:1312.6885 (2013).

[9] Dubrovina, A., Kisilev, P., Ginsburg, B., et al., "Computational mammography using deep neural networks", Comp Methods in Biomechanics and Biomed Engineering: Imaging & Visualization, 1–5 (2016).

[10] Wang, D., Khosla, A., Gargeya, R.., et al., "Deep learning for identifying metastic breast cancer", arXiv (2016).

[11] Cheng, J., Ni, D., Chou, Y., et al. "Computer-Aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in CT scans," Scientific reports, 6 (2016).

[12] Gallego, J., Montoya, D., Quintero, O. "Detection and Diagnosis of Breast Tumors using Deep Convolutional Neural Networks," Conference Proceedings of the XVII Latin American Conference on Automatic Control, 11-17 (2016).

[13] Suzuki, S., Zhang, N., Homma, N., et al. "Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis," IEEE SICE, 1382-1386 (2016).

[14] Wang, D., Khosla, A., Gargeva, R., et al. "Deep learning for identifying metastatic breast cancer," arXiv preprint arXiv:1606.05718 (2016).

[15] Wang, J., Yang, X., Cai, H., et al. "Discrimination of breast cancer with microcalcifications on mammography by deep learning," Scientific reports, 6 (2016).

[16] Kallenberg, M., Petersen, K., Nielsen, M., et al. "Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring," IEEE Trans. Med. Imaging, 35(5) 1322-1331 (2016).

[17] Mohamed, A., Berg, W., Peng, H., Luo, Y., et al., "A deep learning method for classifying mammographic breast density categories", Medical Physics (2017).

[18] Domingues, I., and Cardoso, J.S., "Mass detection on mammogram images: a first assessment of deep learning techniques," 2013.

[19] Dhungel, N., Carneiro, G., Bradley, A.P., "Automated mass detection in mammograms using cascaded deep learning and random forests," in International Conference on Digital Image Computing: Techniques and Applications, IEEE, 1–8 (2015).

[20] Huynh, B.Q., Li, H., and Giger, M.L., "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," Journal of Medical Imaging, 3(3) 034 501– 034 501 (2016).

[21] Levy, D. and Jain, A., "Breast mass classification from mammograms using deep convolutional neural networks," arXiv:1612.00542 (2016).

[22] Krizhevsky, A., Sutskever, I., Hinton, GE., "Imagenet classification with deep convolutional neural networks", Advances in Neural Information Processing Systems, NIPS 25, 1106-1114 (2012).

[23] Heath, M., Bowyer, K., Kopans, D., et al., "The Digital Database for Screening Mammography", Proc. Of 5th Interl. Workshop on Digital Mammography, Physics Publishing, M.J. Yaffe ed., ISBN 1-930524-00-5, 212-218 (2001).

[24] Heath, M., Bowyer, K., Kopans, D., et al., "Current status of the Digital Databse for Screening Mammography", Proc. Of 4th Interl. Workshop on Digital Mammography, Kluwer Academic Publishers, 457-460 (1998).

[25] Sharma, A., "DDSM Utility", GitHub, GitHub repository, https://github.com/trane293/DDSMUtility (2015)

[26] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. "ImageNet: a large-scale hierarchical image database", CVPR (2009).

[27] Jia, Y., Shelhamer, E., Donahue, J., Karavev, S., Long, J., et al. "Caffe: Convolutional Architecture for Fast Feature Embedding", arXiv preprint arXiv:1408.5093 (2014).

[28] Roque AC, Andre TC., "Mammography and computerized decision systems: a review," Acad Sci 980 83–94 (2002).