Automatic identification of nuanced imaging features in recalled but biopsy benign mammogram images

Purpose

Digital 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]. High recall (asking a woman back for additional workup after a screening mammogram) rates is, however, a concern in breast cancer screening. Approximately 10% of women in the U.S. (over 5 million women annually) who are screened using digital mammography are recalled; and 80% of more than 1 million breast biopsies performed annually are benign [1], resulting in unnecessary psychological stress, medical costs, and clinical workload. Improving the interpretation of digital mammograms in clinical reading towards the reduction of high recall rates is of great clinical significance and an unmet need.

Individual radiologists' reading performances in breast cancer detection in screening mammography have been shown to vary widely and have to do with experience, reader volume, and a reader's subspecialty, among other factors. One aspect of improving the mammographic reading is to enhance the radiologists with powerful computerized tools to identify imaging features to aid in clinical decision-making. In order to help radiologists make more accurate decisions on whether to recall a patient, we focus on building computer-aided classifiers that can distinguish the subtle imaging characteristics of the mammograms between patients who were recalled but biopsy benign, patients who were read as negative from the onset, and patients with cancer positive scans. We expect these computerized classifiers can assist radiologists by predicting which patients/images may be falsely recalled but most likely tend to be a benign. Furthermore, the subtle imaging appearance features can be identified and visualized through the classification processes so that can be used to educate radiologists and improve their reading on potential false recalls.

Deep learning [2] coupled with a big training dataset has shown promising performance in many artificial intelligence applications and is entering the field of biomedical imaging. The main architecture of deep learning for image data is the convolutional neural network (CNN) [2]. CNNs are biologically-inspired multi-layer neural networks and each layer consists of connected neurons that have learnable weights. 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 [2]. When labelled clinical data is limited in availability for training a CNN model, deep learning models can be pre-trained with other larger datasets and then fine-tuned using a relatively smaller target dataset.

In this study we investigated how the newly emerged deep learning CNN models can be used for the automatic identification of nuanced imaging features to distinguish mammogram images belonging to negative, recalled-benign, and positive cases, aimed to better interpret the recalled mammographic images with biopsy benign results and ultimately help reduce unnecessary recalls.

Methods

Overview: In order to test the effect of the deep learning models on the classification of digital mammography images, we utilized a combination of two independent mammography datasets (**3923** patients and **15714** images; see below) for training and testing the network. Testing was done in five classification scenarios (four binary-class comparisons plus one triple-class comparison): Positive vs Recalled-Benign+Negative (Recalled-benign and Negative are treated as one class with equal representation within that class), Positive vs Negative, Positive vs Recalled-Benign, Negative vs Recalled-Benign, and Positive vs Recalled-Benign vs Negative (triple-class comparison). 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 maintained a cancer-free status in 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 biopsy-proven benign based on pathology results.

Study cohort and datasets: We used a combination of two independent mammogram datasets, FFDM and DDSM, with a total of **3923** patients and **15714** images in our study. The two datasets were briefly described in the following.

Full-field Digital Mammography (FFDM) Dataset: This is a retrospective cohort of 1303 patients (5234 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; 375 patients were evaluated as positive for breast cancer (27% Ductal carcinoma in situ [DCIS]; 73% Invasive) based on pathology results. Both craniocaudal (CC) and mediolateral oblique (MLO) image views were used for all patients.

Digital Database of Screening Mammography (DDSM) Dataset: The DDSM dataset [3] is a 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. Cases are roughly divided between normal (i.e. negative), cancer (i.e. positive), and benign (i.e. recalled, but determined benign through pathology).

Deep learning approach for building classifiers: Two-class and three-class CNN models were constructed to classify four scenarios of two-category classification outlined above (e.g., "Recalled-Benign vs Negative") and one three-category classification. The CNN used an improved version of the AlexNet model[2], pre-trained by the ImageNet imaging dataset[4]. The weights from the pre-trained model are used as the initial weights in constructing the target CNN models, where these weights are fine-tuned by continuing the backpropagation using the mammography imaging datasets. 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. Key steps in the training process include: (1) Run a histogram equalization on all training images to calibrate the contrast. (2) Resize all training images to 227x227, the standard image size for deep learning on Alexnet for computational efficiency (higher image sizes were tried but performance results showed little variation). (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. The deep learning network was implemented using the Caffe platform running on a 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).

Strategy for evaluating the classifiers: The two datasets were combined (mixed up) to examine (including both training and testing) the CNN-based classification accuracy. The receiver operative characteristic (ROC) curve was generated and the area under the curve (AUC) was calculated as a metric of the classification accuracy. For training, 5028 images were used for training for each of the classification scenarios listed above (2514 images for each category; 867 FFDM images + 1647 DDSM images), except for Negative vs Recalled-Benign (which had more data), in which 8322 images were used for training (4161 for each category; 1520 FFDM images + 2641 DDSM images). For testing, we used 272 images for each scenario (136 for each category, approximately 5%; 50 FFDM images, 86 DDSM images) except in the Negative vs Recalled-Benign scenario, where 438 images were used (219 for each category; 139 DDSM images + 80 FFDM images).

Results

Figure 1 shows results for all five classification scenarios (four binary classification and one triple classification). **Figure 2** shows the corresponding ROC curves for the four binary classification scenarios and **Figure 3** for the triple classification. It should be noted that since ROC is a binary-class evaluation method, we followed the common practice in the literature for triple-class evaluation, i.e., generating an ROC curve for each two-class combination and then reporting the average of the AUCs as the metric of triple-classification assessment.

In all comparisons, our results showed that the three categories (negative, recalled-benign, and positive) can be well distinguished (AUC ranging from 0.76 to 0.91) by automatically identified mammographic imaging nuances. The identified imaging features between recalled-benign and negative are most distinguishing (AUC=0.91). This is also the scenario with the greatest amount of training and testing data. Positive vs Negative (AUC=0.84) had the second best performance, followed by the triple-class comparison (average AUC=0.79). Positive vs Negative+Recalled-Benign and Positive vs Recalled-Benign had similar performance results at AUC=0.76.

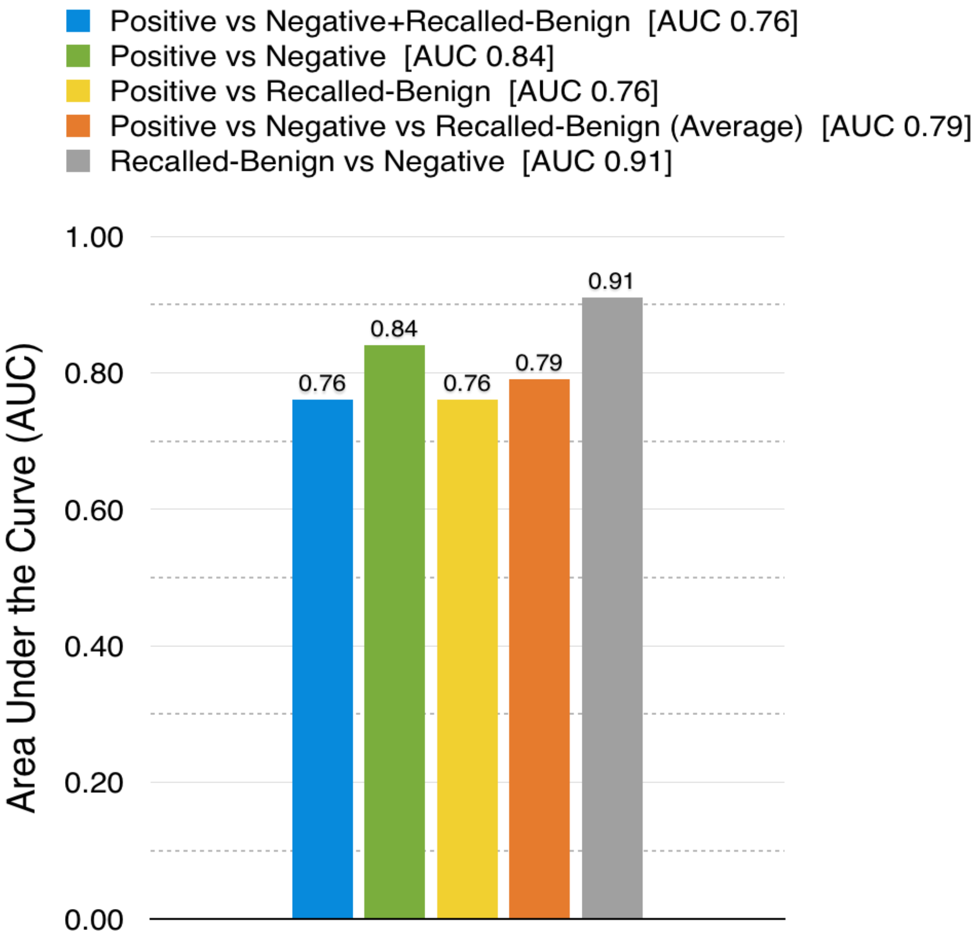

Figure 1. Performance results (area under the receiver operating characteristic curves) for deep learning CNN models for the classification of digital mammography images in five classification scenarios. Note: this represents an updated result from initial submission with extended dataset.

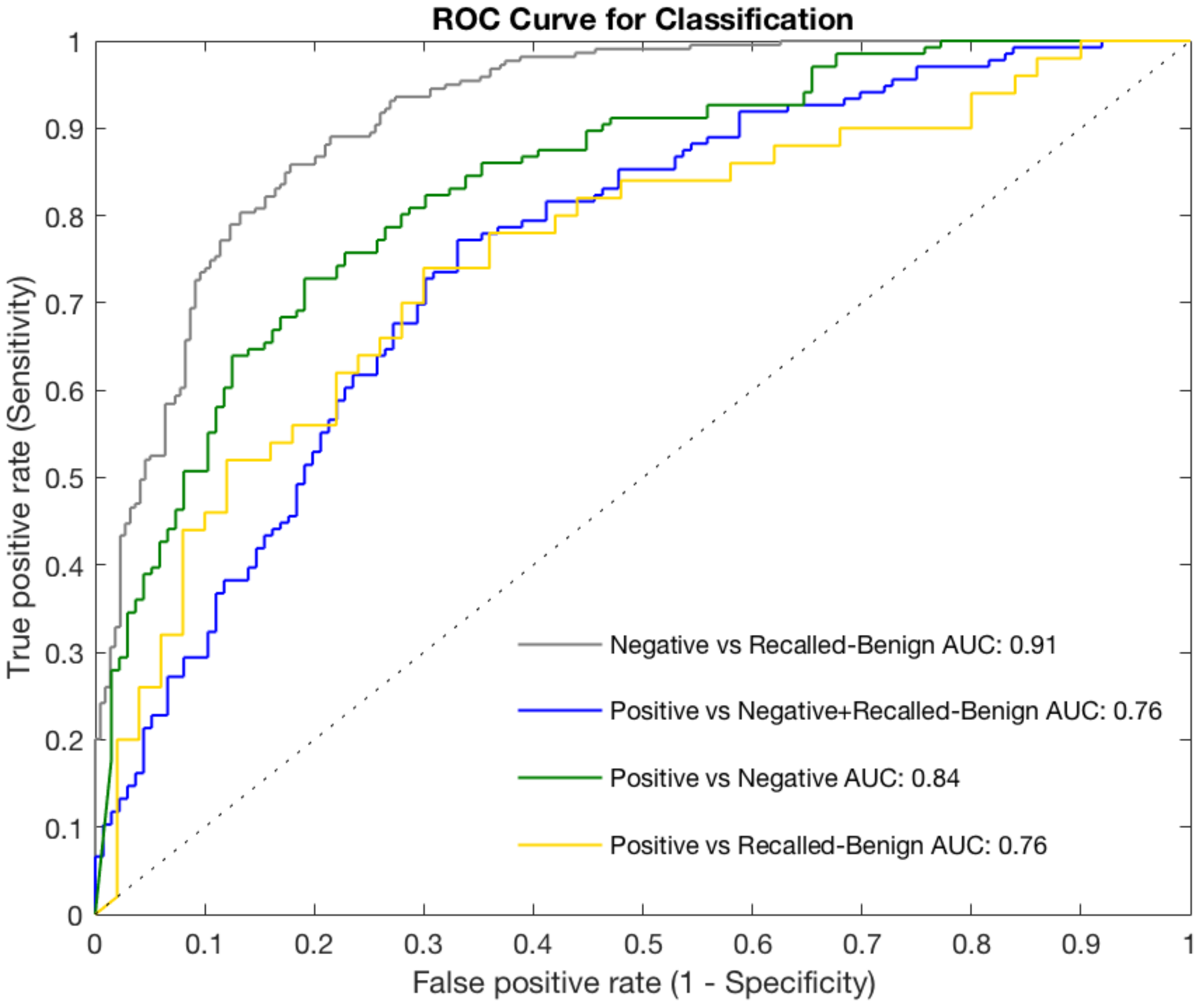

Figure 2. Receiver operating characteristic curves for deep learning CNN models for the four binary classification scenarios.

Conclusions

We presented a novel investigation into the use of deep learning networks for the nuanced imaging feature identification and classification of mammogram images, focusing on five classification scenarios: Positive vs. Negative+Recalled-Benign, Positive vs Negative, Positive vs Recalled-Benign, Positive vs Negative vs Recalled-Benign, and Negative vs Recalled-Benign. In all cases, distinction between the categories was encouraging. The scenario with the greatest data, Negative vs Recalled-Benign showed the best performance. This result, in conjunction with the ability for the model to distinguish recalled-benign and positive images, indicate that there are imaging features unique to recalled-benign images that computerized tools can identify and use to help radiologists make better decisions on whether a patient should actually be recalled or more likely be a false recall. Similarly, with the ability of the CNN model to distinguish between negative and positive images, and all three categories at once in the triple-class classification, the results indicate that automatic deep learning methods can perform well in computer-aided diagnosis of breast cancer. In general, classifiers on these scenarios will contribute to generating intelligent computer-aided clinical tools to help reduce false recalls. We believe our study holds a great potential to incorporate deep learning-based artificial intelligence into clinical workflow to improve radiological reading of mammograms. Additional comprehensive results that could not fit in this abstract due to the word limit will be presented in upcoming journal papers, including comparison of the performance between the two independent datasets and effects of adopting different transfer learning strategies.

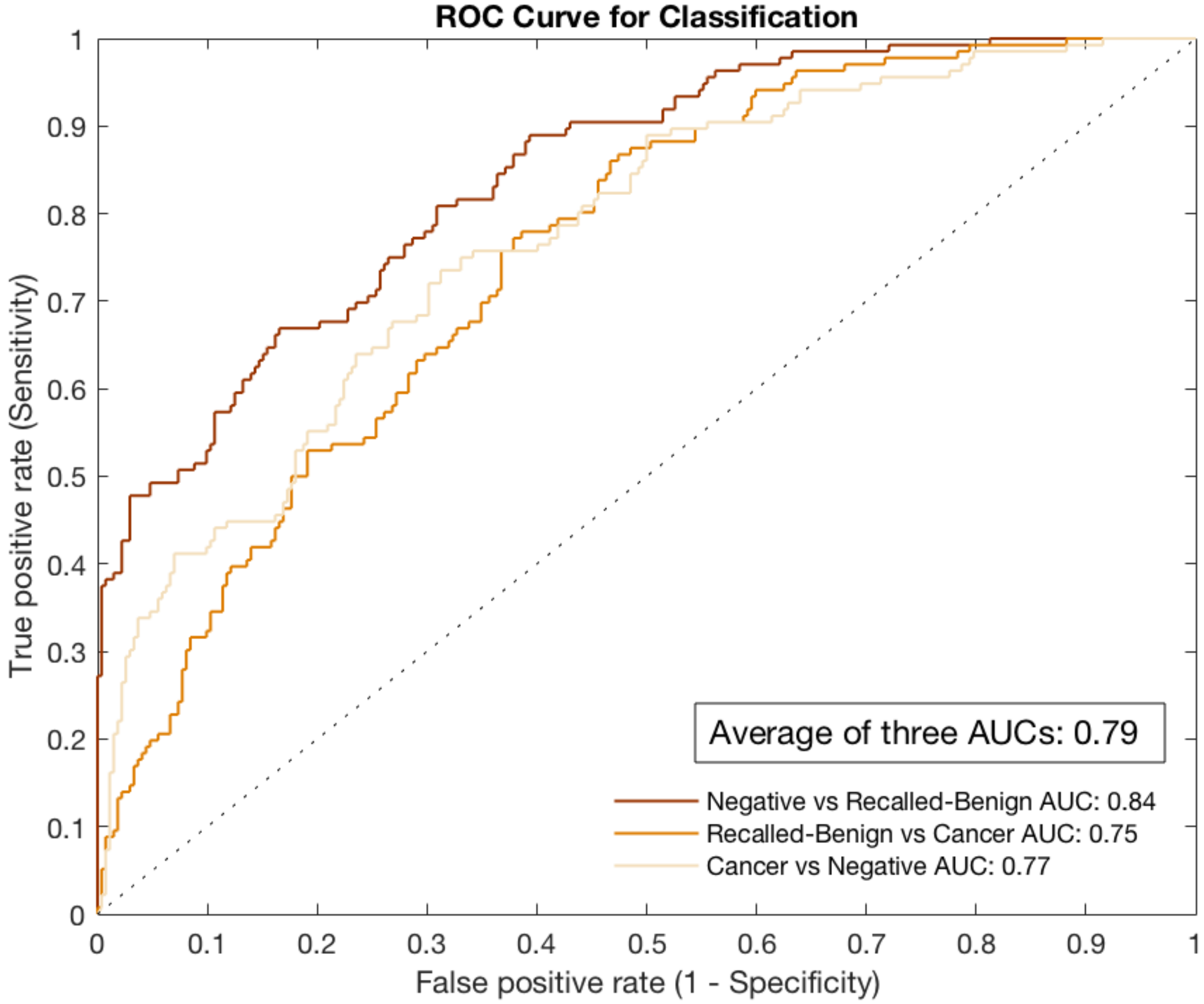

Figure 3. Receiver operating characteristic curves for the triple-class classification scenario and the averaged AUC.

References

[1] 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”,Cancer 106, 732-42 (2006).

[2] LeCun Y., Bengio, Y., and Hinton G., Deep learning, Nature, vol. 521, pp. 436–444, 2015.

[3] 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).

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