DeepCAT: Deep Computer-Aided Triage of Mammograms

Paul H. Yi, MD, Johns Hopkins University School of Medicine; Dhananjay Singh, MS; Susan C. Harvey, MD; Gregory D. Hager, Ph.D; Lisa A. Mullen, MD

Introduction

Although much deep learning research has focused on mammographic detection of breast cancer, relatively little attention has been paid towards mammography triage for radiologist review.

The purpose of this study was to develop and test DeepCAT, a deep learning system for mammography triage based on suspicion of cancer. Specifically, we evaluate DeepCAT’s ability to provide two augmentations to radiologists: 1) discarding images unlikely to contain cancer and 2) prioritization of images likely to contain cancer for radiologist review.

Hypothesis

DeepCAT can facilitate mammography triage by discarding a large proportion of normal studies and prioritizing review of studies likely containing breast cancer.

Methods

We obtained 1878 2D-mammographic images (CC & MLO) from the Digital Database for Screening Mammography (1438 with masses [77%] and 440 normal [23%]); each mass was biopsy-proven to be benign or malignant and annotated with radiologist-performed segmentation masks. Images were randomly divided into training (55%), validation (13.5%) and test splits (31.5%) Table 1 and used to train, validate, and test DeepCAT.

Table 1. Dataset Splits for Development & Testing of DeepCAT

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Malignant</th>
<th>Benign</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training (55%)</td>
<td>1027</td>
<td>490</td>
<td>437</td>
<td>100</td>
</tr>
<tr>
<td>Validation (13.5%)</td>
<td>256</td>
<td>83</td>
<td>73</td>
<td>100</td>
</tr>
<tr>
<td>Testing (31.5%)</td>
<td>595</td>
<td>152</td>
<td>203</td>
<td>240</td>
</tr>
<tr>
<td>Total</td>
<td>1878</td>
<td>725</td>
<td>713</td>
<td>440</td>
</tr>
</tbody>
</table>

DeepCAT (Deep Computer-Aided Triage) is comprised of 2 components: 1) Mammogram classifier cascade and 2) Mass detector, which are combined to generate an overall priority score (described below).

The mammogram classifier cascade consists of 2 ImageNet-pretrained ResNet-34-based classifiers: one to maximize the malignancy recall and non-malignant precision (“malignancy-weighted”) and one to maximize accuracy (“balanced”). The classifier cascade outputs a weighted prediction average for normal vs. benign mass vs. malignant mass from these 2 classifiers.

The mass detector is a RetinaNet-based classifier using ResNet-50 as the Feature Pyramid Network and was trained to classify between background, benign and malignant lesions with bounding boxes generated around potential malignant masses.

DeepCAT prioritizes studies first by using the malignancy-weighted classifier to discard normal studies; because this classifier has high malignant recall, its normal predictions are very precise. All remaining images are then processed through the remainder of DeepCAT with output of classifier cascade weighted prediction average and addition of mass detector prediction for any image with malignant mass probability >50%; this weighted suspicion score is used to order images for radiologist review.
We calculated the theoretical decrease in workload by DeepCAT’s discarding of normal studies. We then measured prioritization performance by computing the number of adjacent swaps necessary to reorder images to an optimal study review order by which malignant images are prioritized over non-malignant ones (“Priority Ordinal Distance” [POD]).

Results
Of 595 testing images, DeepCAT recommended discarding 315 images (53%), of which none contained a malignant mass. Of the remaining 280 images not discarded, DeepCAT’s study ordering required a POD of 25 swaps to obtain perfect prioritization order.

Conclusion
Our results suggest that DeepCAT could substantially reduce a mammographer’s workload by approximately ½ and effectively triage review of mammograms with malignant masses over benign ones. Next steps include application of DeepCAT to digital mammography (2D & tomosynthesis) and prospective clinical validation.

Statement of Impact
If clinically deployed, DeepCAT could substantially reduce a radiologist’s workload and reallocate radiologist attention to images that would most benefit from review for potential breast cancer.

Keywords
breast cancer, mammography, screening, deep learning, radiologist augmentation