Automated Breast Cancer Diagnosis Using Deep Learning and Region of Interest Detection (BC-DROID)

Richard Platania*
Louisiana State University
Baton Rouge, Louisiana 70803
rplata1@lsu.edu

Shayan Shams*
Louisiana State University
Baton Rouge, Louisiana 70803
sshams2@cct.lsu.edu

Seungwon Yang
Louisiana State University
Baton Rouge, Louisiana 70803
seungwonyang@lsu.edu

Jian Zhang
Louisiana State University
Baton Rouge, Louisiana 70803
zhang@csc.lsu.edu

Kisung Lee
Louisiana State University
Baton Rouge, Louisiana 70803
lee@csc.lsu.edu

Seung-Jong Park
Louisiana State University
Baton Rouge, Louisiana 70803
sjpark@cct.lsu.edu

ABSTRACT

Detection of suspicious regions in mammogram images and the subsequent diagnosis of these regions remains a challenging problem in the medical world. There still exists an alarming rate of misdiagnosis of breast cancer. This results in both over treatment through incorrect positive diagnosis of cancer and under treatment through overlooked cancerous masses. Convolutional neural networks have shown strong applicability to various image datasets, enabling detailed features to be learned from the data and, as a result, the ability to classify these images at extremely low error rates. In order to overcome the difficulty in diagnosing breast cancer from mammogram images, we propose our framework for automated breast cancer detection and diagnosis, called BC-DROID, which provides automated region of interest detection and diagnosis using convolutional neural networks. BC-DROID first pretrains based on physician-defined regions of interest in mammogram images. It then trains based on the full mammogram image. The resulting network is able to detect and classify regions of interest as cancerous or benign in one step. We demonstrate the accuracy of our framework's ability to both locate the regions of interest as well as diagnose them. Our framework achieves a detection accuracy of up to 90% and a classification accuracy of 93.5% (AUC of 92.315%). To the best of our knowledge, this is the first work enabling both automated detection and diagnosis of these areas in one step from full mammogram images. Using our framework's website, a user can upload a single mammogram image, visualize suspicious regions, and receive the automated diagnoses of these regions.

KEYWORDS

mammogram; breast cancer; deep learning; automated diagnosis; convolutional neural network; object detection

1 INTRODUCTION

Breast cancer is the most common and fatal cancer among adult women [36]. According to National Cancer Institute, approximately one in eight women will develop an invasive form of this cancer at some point in their lives [26]. Frequent screenings through mammograms can help detect early signs of breast cancer. However, there is still difficulty in recognizing troublesome areas, and applying a correct diagnosis, through these images alone. For example, breast cancer overdiagnosis is estimated to be 22~31% of all diagnosed breast cancers, costing several billions of dollars annually in healthcare spending [28]. In order to alleviate the misdiagnosis rate of mammogram screenings, further image analysis must be performed in order to provide physicians with an aid in diagnosis.

Existing methods for breast cancer detection and diagnosis have improved upon physician-only detection and diagnosis. Current works can be divided into two categories. The first focuses upon mass detection or segmentation in mammogram images. This aids the physician by pinpointing potential suspicious regions in the mammogram that may have been overlooked. While typical computer-aided detection (CAD) has shown little to no improvement in this category [22], more advanced machine learning and deep learning techniques have shown promise towards the detection and segmentation tasks [7–10, 17, 29]. The second category aims to diagnose breast cancer from mammogram images (or the masses). Some works have utilized more traditional machine learning methods for this task [38], while others have moved towards deep learning [3–5, 11, 23]. However, existing works have failed to combine these two tasks into one tool that simultaneously provides region of interest detection and diagnosis of mammogram images with visualization.

In order to further automate and improve upon the detection and diagnosis of breast cancer, we propose an automated diagnosis tool for mammograms, called BC-DROID, that uses deep learning for two outcomes; it detects and localizes any regions containing abnormalities and provides a diagnosis based on these regions. A simplified workflow diagram of BC-DROID is given in Figure 1. Given any mammogram image, our tool will provide visualization of regions of interest and the resulting diagnosis. As a result, our unique contribution consists of using a single model to both detect regions of interest and diagnose them, from complete mammogram

*These authors contributed equally to this work.

ACM-BCB'17, August 20-23, 2017, Boston, MA, USA.
© 2017 ACM. 978-1-4503-4722-8/17/08...$15.00
DOI: http://dx.doi.org/10.1145/3107411.3107484

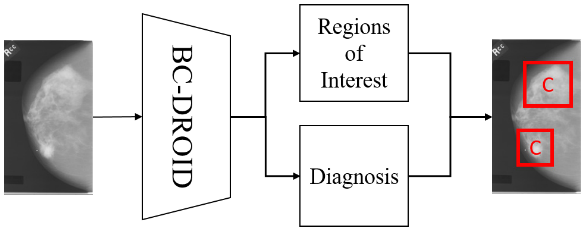


Figure 1: A diagram depicting the process of BC-DROID. Given a single mammogram image, it will detect regions of interest and diagnose them as cancer or benign.

images. The resulting model is made available through an efficient, easy-to-use website.

The remainder of the paper is organized as follows. First, we give some background of breast cancer screening with mammography and convolutional neural networks for object detection and classification. Next, we describe related works. Following, we describe our methodology, which includes details regarding our dataset, convolutional neural network model, and automated diagnosis tool. We then give results showing the accuracy of our work and its ability to detect and diagnose abnormalities in mammograms. Finally, we conclude and discuss future works.

2 BACKGROUND

2.1 Breast Cancer Screening with Mammography

Since its inception in 1913, mammography has been used for early detection and diagnosis of breast cancer by radiologists. It gained a wide acceptance by incorporating general purpose x-ray tubes, films, and techniques such as breast compression and use of fine-grain intensifying screens in the 1950s. Follow-up studies during this time showed a significant reduction in breast cancer mortality rate by almost one-third of women who had mammography screening [14, 37]. Mammography techniques improved significantly with the introduction of screen-film mammography (SFM) and magnification mammography in the 1970s. SFM, which is the gold standard for breast cancer detection, allowed rapid processing, shorter exposure to radiation, and sharper high contrast images for increased 'see-through' examination. FDA approved digital mammography (DM) systems in 2000 for efficient storing and analysis of mammograms using computers. It is reported that the overall diagnostic accuracy between SFM and DM was similar. However, DM showed significantly higher accuracy in women under 50 with dense breasts, and lower recall rate than SFM [24, 30].

Other screening technologies such as untrasound and breast magnetic resonance imaging (MRI) have been adopted in the 2000s as adjunctive screening tools along with a traditional mammography for women who have higher risk of developing breast cancer. Ultrasound screening was effective in women who have dense breast tissue and negative mammograms. MRI screening showed higher sensitivity for women with high-risk of developing hereditary breast cancer [19, 21]. However, it was also reported that both technologies-untrasonic and MRI-yielded high false-positive rates as well [21].

Recently, the 3D mammography technology, also known as digital breast tomosynthesis, emerged. It facilitated layer-by-layer examination of the breast, which enabled viewing of fine details that might have been unrecognizable in traditional mammograms. Further, the integrated 2D and 3D mammography increased cancer detection rate by 51% across all ages compared to that of 2D mammography, at the same time reducing false positive recalls [6, 16]. Multiple institutions such as Society of Breast Imaging, American Cancer Society, and American College of Radiology recommend a periodic screening for breast cancer (e.g., annual screening from age 40 for women at average risk, and age between 25 and 30 for women with BRCA1 or BRCA2 mutation carriers or with family history of breast cancer). However, controversy exists about the starting age of periodic mammography [18, 21].

2.2 Convolutional Neural Networks

Convolutional neural networks begin with (typically) at least several convolution layers and end with one or more fully connected layers. Between the convolution layers often lies pooling layers that perform subsampling on the data to reduce training overheads. Sometimes, between the convolution layers, there are also normalization layers, but they have not seen significant usage in recently developed models. The entire network will take as input an image of size (h,w,c), where h is the height, w is the width, and c is the number of channels in the image. These channels typically refer to different colors (RGB), so typically c has a value of three. Each convolution layer has many filters, the size of which is smaller than the input, that independently perform the convolutions across the image. These filters learn patterns across the entire image. As the input is passed through the network, the convolution layers perform convolutions on the image.

Because CNNs assume that the input is an image, they have several important structural changes compared to a traditional neural network. Neurons in neighboring layers exhibit a local connectivity pattern, ensuring that filters learn to detect patterns based on spatial locality. Many filters are stacked, producing a 3D volume of neurons that is capable of detecting many different patterns. Each filter scans the entire image for patterns. However, in order for a filter to detect a pattern across the whole image, weight sharing is used. This ensures that the model does not need to individually learn to detect a certain pattern at different positions in the image. Instead, a filter learns a certain pattern that it can detect across the whole image.

3 RELATED WORKS

3.1 Mammogram-based Classification Tasks

Deep learning has shown substantial applicability to medical image analysis in recent years. Furthermore, many of these efforts have been towards utilizing mammography data for classification tasks. Peterson et. al. and Kallenberg et. al. utilized unlabeled mammogram images and unsupervised deep learning for the classification of breast tissue segmentation, percentage mammographic density (PMD) score, and mammographic texture (MT) score [17, 29]. They employ a convolutional neural network (CNN) and sparse autoencoder for these tasks. As a result of this unsupervised learning, they aimed to better determine future cancer risk in patients. Other similar works have made use of CNNs for classification of regions of interest or mass lesions in mammograms. Arevalo et. al. use supervised learning with CNNs for representation learning and classification of breast mass lesions as benign or malignant [3, 4]. In a similar manner, Lévy et. al. trained a CNN for classification of breast mass lesions [23]. Abbas also utilizes deep learning for the classification of breast mass from predefined regions of interest [2]. Some works still make use of more traditional methods of image analysis, such as the work of Zhang et. al., which uses Fourier transforms and principal component analysis, followed by a support vector machine (SVM), to classify regions of interest [38]. While these works focus on classifying cancer risk based on full images, they do not detect and localize masses in the image. However, there exists the issue that these works rely on a predetermined region of interest, whereas our work is capable of processing the entire image with automated extraction of these regions. Two additional works leverage multiple image views, as seen in Figure 2, by training multiple CNNs (one for each view). Carreiro et. al. uses a CNN that has been pretrained on the ImageNet dataset [33] to estimate the risk of a patient developing breast cancer[5]. Since the images contained in the ImageNet dataset are vastly different than those in mammogram datsets, it is better to pretrain a model based on mammogram data. Our work pretrains the CNN based on regions of interest. Similar to the previously mentioned work, Geras et. al. classify based on risk, but instead use high-resolution images the network is not pretrained [11]. Our work, unlike these multi-view mammogram works, does not require multiple views of the breast for performing detection and classification.

3.2 Region of Interest Detection in Mammograms

Region of interest detection requires using the entire mammogram image. Processing the entire image is a much more challenging process. While the regions of interest may only be a few hundred pixels on any side, the whole image tends to have thousands of pixels. However, it is important to consider these full images since, in order to fully automate the diagnosis process, the regions of interest should be required to be predefined by an expert source (i.e., a physician). To date, there are a limited number of deep learning related works that utilize the entire mammogram image. Dhungel et. al. extended their previously mentioned works to include a deep belief network (DBM) capable of generating candidate regions of interest [7]. Unlike our work, which only looks at the image once, they require processing multiple scales of the same image, which further increases the complexity and time involved in analysis. Furthermore, their work focuses solely on the detection of masses and ignores the classification of these masses as cancerous or benign. Also operating on the full image, Ertosum et. al. train two CNNs [10]. The first CNN classifies a mammogram as containing or not containing a mass. Following this, the second CNN identifies and localizes the mass. Similar to the previously mentioned work, they do not classify them as malignant or benign. Because of the computational challenges involved, works related to region of interest detection are limited. Unlike the other works, BC-DROID is able to detect regions of interest and classify them from full mammogram images.

3.3 Convolutional Neural Networks for Object Detection and Classification

Convolutional neural networks (CNN) have played a major role in both object detection and classification of image data. Their popularity has sparked an influx in their applicability to many domains. Additionally, competitions, such as ILSVRC [33], push deep learning experts towards development of newer, more impressive CNN models [20, 34, 35]. Of particular difficulty is object detection. In general, it requires more time and complexity to train a model for this task when compared to classification. This is because techniques for this task tend to revisit the image multiple times or explore multiple different scales of the image. Several works have been created with the goal of reducing the time and complexity of object detection [12, 13, 25, 31, 32].

For this work, we make use of the techniques described in the YOLO paper[31]. YOLO allows for object detection and classification while processing the image only once, as is implied by it's name, You Only Look Once. This significantly reduces the time required for our task. Since we are dealing with processing large-scale images through our web-based tool, speed is important.

4 METHODS

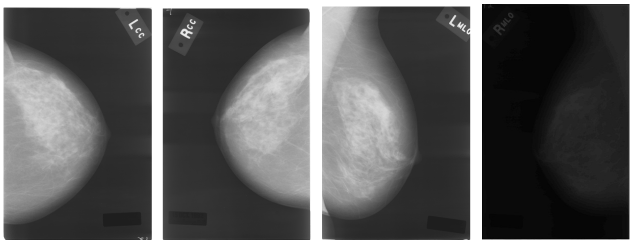
4.1 Mammogram Data

This work uses the Image Retrieval in Medical Applications (IRMA) version of the Digital Database for Screening Mammography (DDSM) dataset, as well as some additional metadata provided directly by the original DDSM dataset [15, 27]. There are 2,620 cases, divided into three diagnoses: normal, benign, and cancer. Each case includes an overview file, four mammogram images, and zero to four overlay files. The overview file provides information such as the date of the experiment, age of the patient, and list of the image files. Each overlay file is associated with one image. It describes the number of abnormalities found in the image as well as the shape and location of the abnormalities. If there are no abnormalities in an image, there is no associated overlay file. Examples of the mammogram images are provided in Figure 2. It is worth noting the low quality of some images, as is visible in Figure 2(d). In total, there are 10,480 images, many of which have a size around 3,000 x 5,000 pixels. For further information regarding the DDSM dataset, see the official website[1].

4.2 Automated Regions-of-Interest Detection and Diagnosis using Mammogram

4.2.1 Data preprocessing. There were several steps taken to preprocess data and improve the training process. In order to remove unnecessary noise, such as the words visible in Figure 2, much of the black portion of the images was trimmed. This was done by flipping all right view images to left and trimming from right to left while the mean pixel value remained near zero (black). Several more common approaches were taken to reduce overfitting the model and improve overall accuracy. For example, each image was rotated five times at random angles and randomly mirrored across the y-axis. Following this, for n pixels, each pixel p_i was normalized

[1] http://marathon.csee.usf.edu/Mammography/Database.html



(a) LEFT_CC (b) RIGHT_CC (c) LEFT_MLO (d) RIGHT_MLO

Figure 2: Sample mammogram data from DDSM showing four different views of breast. Each figure may have noise as seen by the writing in the corners. Some images are also of low contrast, such as that seen in (d).

as follows:

$$p_i = p_i - \frac{\sum_{j=0}^{n} p_j}{n} \tag{1}$$

Note that it is important that the majority of unimportant, dark pixels were removed from the images prior to this step. Otherwise, they drastically affected the image mean and overall reduced the image quality.

Further processing was required for extracting the regions of interest from the mammograms for pretraining our model. To this end, the overlay files provided with the DDSM dataset were used. Since the regions of interest defined in these files take many shapes, we cropped the region such that the crop was the smallest possible square that contained the entire region. The extracted regions were then resized to 128 × 128 for pretraining.

4.2.2 *Model for Region-of-Interest Detection and Diagnosis.* We adapt the object-detection model proposed in YOLO [31] to identify region of interest (ROI) (and label the region as benign or cancer) for mammogram images. Our adapted model takes a mammogram image as input and outputs multiple predictions. In particular, we impose a $k \times k$ grid on the image. For each grid cell, the model predicts the following:

- A confidence value (c). It indicates the confidence that a region of interest (ROI) exists with respect to the grid cell. Similar to YOLO, we define the confidence to be the probability that the grid cell contains the center of a ROI, multiplied by the IOU (intersection over union) ratio of the ROI and the grid cell area.
- The coordinates (x and y) of the center of the ROI, relative to the grid cell. A grid cell is only responsible for the ROI whose center is inside the grid cell. Hence, $0 \le x \le 1$ and $0 \le y \le 1$.
- The width (w) and the height (h) of the ROI, relative to the size of the image.
- A class label vector (P). Because we consider diagnosis between two classes: benign and cancerous, the vector is of length 2, each element representing the probability of the corresponding class.

In summary, the model makes 7 predictions $(c, x, y, w, h, P[0], P[1])$ per grid cell. The total number of values predicted by the model is $7k^2$. Note that, the prediction for each of the values is based on the whole image, not the part of the image in the corresponding grid cell.

The deep neural network model for making these predictions has two main parts. The first part is a stack of 3 × 3 convolution layers, of which the first six are followed by 2 × 2 pooling layers. Convolution layers function as feature extractors that generate a high-level representation of the mammogram image and pooling layers provide invariance to small sequence shifts to the left or right and reduce the dimension of the input to the next layer. This CNN and its configuration are shown on the left in Figure 3. The second part is the last three fully connected layers. The last layer has $7k^2$ neurons and its outputs are the predictions described above. The full model is shown on the right side of Figure 3.

4.2.3 *Pretraining CNN on regions of interest .* After preprocessing the mammogram images, data augmentation and extracting the regions of interest, we generate 25,000 cropping of the area of interests and use them to pretrain the CNN part of the model. The pretraining process uses a modified version of the model, where the last fully connected layer is changed to have only two neurons. The modified model is trained to classify the cropping of the area of interests into two classes: benign or cancerous. After pretraining, the learned weights for the CNN part of the model are extracted and will be used as the initialization values for the CNN part in the main training process.

4.2.4 *Training model on entire image.* If a grid cell contains the center of a ROI, we say that the ROI is *presented* at that grid cell. Given an image in the training set, for each grid cell that has a ROI presented, we can calculate the true values for the predictions. Denote by $(\hat{c}, \hat{x}, \hat{y}, \hat{w}, \hat{h}, \hat{P})$ these true values. The goal of the training is to fit the predictions to the true values. Similar to YOLO [31], we use squared error in the fitting and for each grid cell i, we have the following loss functions:

$$\mathcal{L}^i_{coo} = \begin{cases} (x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2, & \text{ROI present} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$\mathcal{L}^i_{box} = \begin{cases} (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2, & \text{ROI present} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$\mathcal{L}^i_{cls} = \begin{cases} \sum_j (P_i[j] - \hat{P}_i[j])^2, & \text{ROI present} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

These loss functions count prediction error only if a ROI is presented in the grid cell. Hence, the corresponding predictions are trained by the ROIs presented at that grid cell in the training dataset. If in a training example, the grid cell has no ROI, this example won't have any effect for training the above predictions for the cell. On the other hand, the empty case, i.e., grid cell with no ROI presented does affect the training of the confidence prediction for the cell:

$$\mathcal{L}^i_{cnf} = \begin{cases} (c_i - \hat{c}_i)^2, & \text{ROI present} \\ \beta(c_i - \hat{c}_i)^2, & \text{otherwise} \end{cases} \tag{5}$$

If there is a ROI at the grid cell i, $\hat{c}_i$ = IOU ratio of the ROI and the cell. If no ROI is presented, $\hat{c}_i = 0$. For any grid cell, there are many more

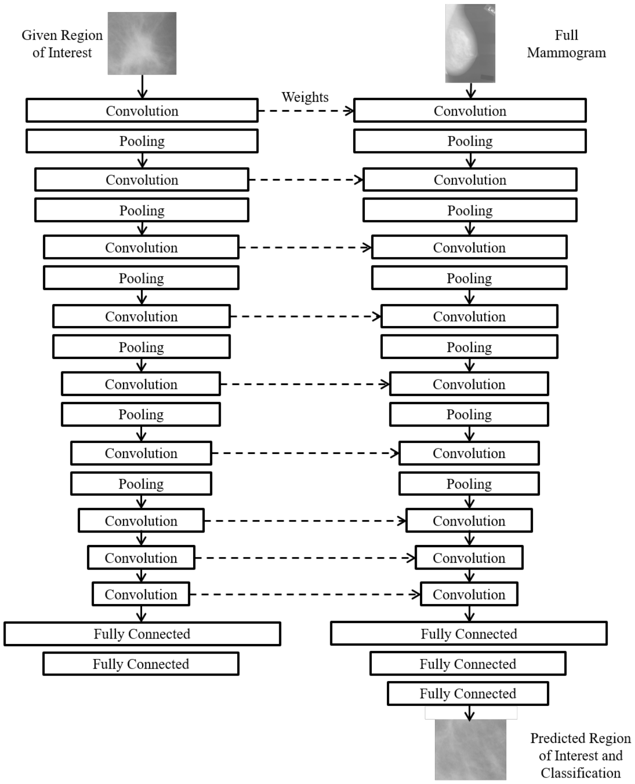


Figure 3: The left represents the pretraining of the model on the regions of interest. The weights from the convolutional layers are then used to initialize the model for training on the whole image.

training examples without ROI than examples with ROI. $0 < \beta < 1$ is used to balance the two cases in training.

The total loss sums up the above losses for each grid cell and then sums over all the grid cells:

$$\sum_{i=1}^{k^2} \alpha(\mathcal{L}^i_{coo} + \mathcal{L}^i_{box}) + \mathcal{L}^i_{cls} + \mathcal{L}^i_{cnf} \tag{6}$$

where α is a parameter similar to β that balances errors between location regression and the others. For our experiments, we found the best values for α and β to be 5 and 0.4, respectively.

To train the model, we minimize the total loss defined above by adjusting the neural network weights. The training uses whole mammogram images, not the ROI cropping. Stochastic gradient descent (SGD) is utilized for the minimization. The weights of the CNN part are initialized by the weights from pretraining. Figure 3 depicts the process of extracting the weights from the pretraining step and using them to initialize the CNN part in the main training. It is worth noting that the weights of the fully connected layers are initialized randomly, not by transferring those from the fully connected layers in the pretrained model.

4.2.5 Inference. To locate the ROI for a testing image, the model is applied to the image to obtain the prediction values (c, x, y, w, h, P) for each grid cell. Given a threshold τ, if $\max cP > \tau$ at a grid cell, then there is a ROI with location and size given by (x, y, w, h), and the class/diagnosis of the ROI is given by $\arg\max_i cP[i]$.

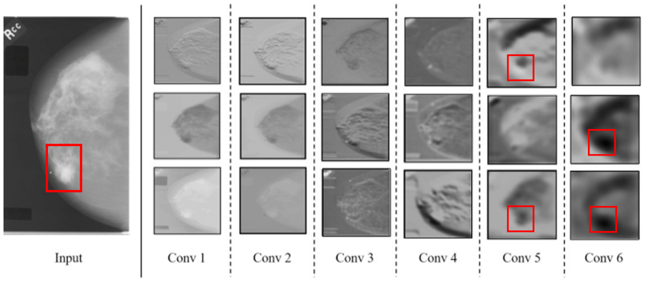


Figure 4: Activations of the first six convolution layers with respect to the given input image. The red rectangle indicates a region of interest. We can observe the convolution layers responding strongly to the region of interest, especially in the later convolution layers (Conv 5 and Conv 6).

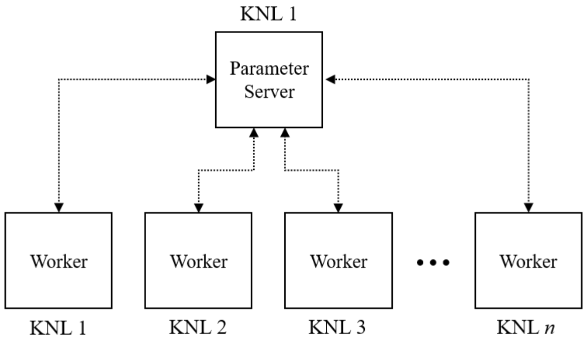
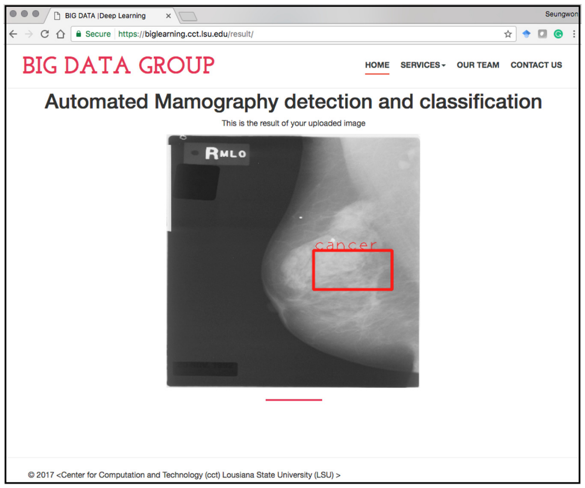


Figure 5: Our training architecture. We define one parameter server on the first Knight's Landing (KNL) node. More workers are defined based on the total number of KNL nodes.

4.2.6 Hardware and software configuration. We used TensorFlow version one [1] for developing our model. Training was distributed among four Intel's Knight landings hosts (Intel Xeon Phi Processor 7230F (16GB, 1.30 GHz, 64 cores)) with one parameter server and four workers. An illustrated depiction of this setup is given in Figure 5. The batch size is 16 and the learning rate was increased from 10^{-4} to 10^{-3} for the first 20 epochs and then with 10^{-3} until 80 epochs. After that, we started decreasing the learning rate to 10^{-4} and then 10^{-5}.

4.2.7 Diagnosis of mammogram image through web-based application. In an effort to make our tool easily available and easy to use, it is available through our website[2]. An example screenshot of the site operating on a mammogram image is given in Figure 6. Since one of the aspect of this work is the applicability as a good assistant to physicians and radiologists, the network is designed to be fast even on a desktop. As a result, our network can process and infer a full mammogram image in 1.25 seconds on a desktop

[2] https://biglearning.cct.lsu.edu/



Automated Mamography detection and classification

This is the result of your uploaded image

© 2017 <Center for Computation and Technology (cct) Lousiana State University (LSU) >

Figure 6: Screenshot of our website providing automated detection and diagnosis of mammogram image.

with 4 cores CPU and 16 GB of RAM, however since one goal is supporting data streaming, the model can scale up and scale out on GPUs. If there is a GPU available on the machine, our model will send the matrix and vector calculation to the available GPUs on the machine. It also is capable of having different parameter server and workers to scale out on different nodes. As a result, our network supports data streaming. The link is available, and the website will soon be available to the public, enabling researchers to stream their images and get the detection and diagnosis results in a very short time, since BC-DROID is supported by HPC platforms with multiple nodes and GPUs.

5 RESULTS

We present our results in two sections. First, we evaluate BC-DROID with respect to correctly detecting regions of interest in mammogram images. Then, we highlight the classification accuracy of these abnormal regions as cancerous or benign. Since to the best of our knowledge we are the first group doing mass classification and region of interest detection in one process, it is difficult to directly compare our detection results with previous works. Thus, for a fair comparison, we divide our result into two sections, detection and classification. We directly compare the classification results with other works. Since some of these works are using BIRADS score or using different data set, we chose the best accuracy score that they obtained for classification.

5.1 Region of Interest Detection

To measure how well our tool detects regions of interest in the images, we first define our accuracy metric. For detection accuracy we use the intersection of union (IOU) or The Jaccard index which is the percentage of overlap between our predicted region of interest and the actual region of interest. Given overlap IOU, which is the percentage of overlap between our predicted region of interest and the actual region of interest, and a threshold t, which defines the required value of o for correct detection, we can compute our accuracy a as follows:

$$a = \begin{cases} 1, & \text{if } IOU \geq t \\ IOU, & \text{otherwise} \end{cases} \tag{7}$$

Table 1: Region of Interest detection accuracy with varying overlap percentage thresholds.

Overlap Threshold t	100%	75%	50%	25%
Accuracy a	53%	56%	84%	90%

Table 2: Classification results.

Specificity	Sensitivity	Accuracy	Precision	Recall
93%	94%	93.5%	93.39%	93%

In other words, if the percentage of overlap between the actual and predicted abnormal regions is above the defined threshold, we correctly detected the region for that particular threshold. Otherwise, if it is below the threshold, we used the calculated IOU score as accuracy score. Table 1 shows the accuracy results using this evaluation metric. When the required overlap between the actual and predicted regions is 100%, our tool has an accuracy of 53%. This score improves as we decrease the threshold. At 25% threshold, we get an accuracy of 90%. In practice, a threshold as low as 25% would still be indicative of the abnormal region in question.

For visualization of this region of interest detection process, we provide Figure 4. This displays the first six convolution layer activations with respect to an example input image. As expected, the earlier layers are very representative of the input image. The activations in the later layers, specifically the sixth, become much more sparse and localized. These localized activations become representative of the regions of interest in the original input. This process depicts our model properly learning to identify regions of interest. As it is shown in Figure 4, deeper CNN layers are concentrating on picking up features representing the region of interest. For better visualization we trace the region of interest for the input image.

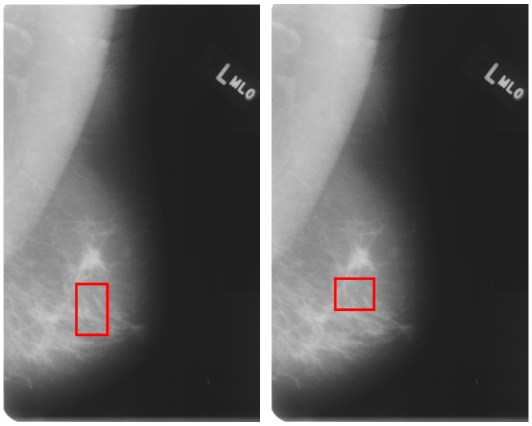
We provide a further visualization in Figure 7. The provided mammogram image is an example of a cancer case. The left image depicts the region of interest as defined by the DDSM dataset. Note that the actual region of interest is not a rectangular shape. Instead, a rectangular box was drawn around the region such that it was minimal and fit the entire region. The image on the right shows our predicted region of interest. Both regions of interest (the actual and predicted) are indicative of the same region in the mammogram image.

5.2 Breast Cancer Classification

Here, we present our results concerning breast cancer classification. Table 2 summarizes our accuracy related results. Additionally, we provide Figure 8, which plots our ROC curve for breast cancer classification. The curve was drawn according to 100 different thresholds for classification. The AUC associated with this curve is

Table 3: Classification comparison with other works using AUC. Their best reported AUC for classification tasks using mammograms was used. If AUC was unavailable, accuracy is used instead.

Paper	Classification Task	Best Reported Result
[17]	Mammographic Texture	61% AUC
	Mammographic Density	62% AUC
[10]	Mass or No Mass	85% Accuracy
[29]	Mammographic Texture	70% AUC
	Breast Cancer Diagnosis	60% AUC
[3, 4]	Breast Cancer Diagnosis	86% AUC
[11]	Breast Cancer Diagnosis	76.5% AUC
[23]	Breast Cancer Diagnosis	92.9% Accuracy
[2]	Breast Cancer Diagnosis	91% AUC
BC-DROID	Breast Cancer Diagnosis	92.315% AUC



(a) Actual

(b) Predicted

Figure 7: A comparison between the actual region of interest, as dictated by a physician, and the region of interest predicted by our automated diagnosis tool.

92.315%. To our best knowledge, these results are the highest of any similar works relating to classification of mammograms as cancer or benign.

We compare our results with related works in Table 3. We reported other works' AUC if available. Otherwise, we substituted in classification accuracy. Since many works differ in classification task, we briefly describe the task of each work in the second column. Tasks included either classifying mammographic texture score (MT), classifying mammographic density (MD) score, determining if an image contains a mass, or diagnosing breast cancer. Of these works, BC-DROID reports the best AUC (and accuracy when compared to works with no AUC).

6 CONCLUSION AND FUTURE WORKS

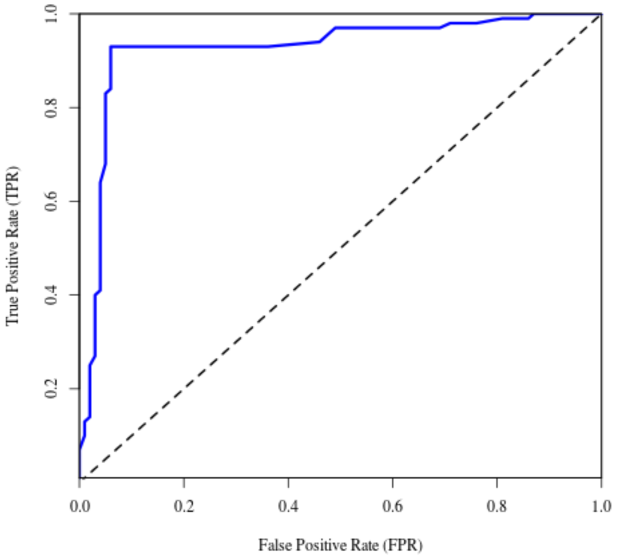


Figure 8: The ROC curve for our classification results. The associated AUC is 92.315%.

In this work, we presented our tool, BC-DROID, for automated detection and diagnosis of breast cancer using mammogram images advanced deep learning techniques. We demonstrated the effectiveness of BC-DROID's detection of suspicious areas and further diagnosis as benign or malignant compared to related works. Our tool is made available through an easy to use web interface, allowing users to upload mammogram images and visualize areas of interest in the image, along with their associated diagnoses. Through this work, we provide physicians with an automated diagnosis that can assist them in making a more conclusive diagnosis. By doing so, this can reduce the rate of false positive diagnoses. Furthermore, it can pinpoint previously unnoticed worrisome areas, lowering the rate of undiagnosed breast cancer. In future work, we are going to frequently update the website with more accurate and faster versions of the pipeline to enable researchers and physicians to have better detection and diagnosis assistance. We plan to migrate our service to an improved High Performance Computing (HPC) system using multiple GPUs and computation nodes in order to enable BC-DROID as a streaming service. This will make it possible to process hundreds of images simultaneously and in a very short period of time. In addition, we plan to extend our technique to other medical imaging data, such as X-ray, fMRI, or others.

ACKNOWLEDGMENTS

This work was partially funded by NIH grants (P20GM103458-10, P30GM110760-03, P20GM103424), NSF grants (MRI-1338051, CC-NIE-1341008, IBSS-L-1620451), LA Board of Regents grants (LEQSF(2016-19)-RD-A-08 and ITRS), and IBM faculty awards.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and

others. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Qaisar Abbas. 2016. DeepCAD: A Computer-Aided Diagnosis System for Mammographic Masses Using Deep Invariant Features. *Computers* 5, 4 (2016), 28.

[3] J. Arevalo, F. A. GonzÃąlez, R. Ramos-PollÃąn, J. L. Oliveira, and M. A. Guevara Lopez. 2015. Convolutional neural networks for mammography mass lesion classification. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 797–800. DOI:https://doi.org/10.1109/EMBC.2015.7318482

[4] John Arevalo, Fabio A. GonzÃąlez, RaÃžl Ramos-PollÃąn, Jose L. Oliveira, and Miguel Angel Guevara Lopez. 2016. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine* 127 (2016), 248 – 257. DOI:https://doi.org/10.1016/j.cmpb.2015.12.014

[5] Gustavo Carneiro, Jacinto Nascimento, and Andrew P. Bradley. 2015. *Unregistered Multiview Mammogram Analysis with Pre-trained Deep Learning Models.* Springer International Publishing, Cham, 652–660. DOI:https://doi.org/10.1007/978-3-319-24574-4_78

[6] Stefano Ciatto, Nehmat Houssami, Daniela Bernardi, Francesca Caumo, Marco Pellegrini, Silvia Brunelli, Paola Tuttobene, Paola Bricolo, Carmine FantÃš, Marvi Valentini, Stefania Montemezzi, and Petra Macaskill. 2013. Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study. *The Lancet Oncology* 14, 7 (2013), 583 – 589. DOI:https://doi.org/10.1016/S1470-2045(13)70134-7

[7] N. Dhungel, G. Carneiro, and A. P. Bradley. 2015. Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests. In *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. 1–8. DOI:https://doi.org/10.1109/DICTA.2015.7371234

[8] Neeraj Dhungel, Gustavo Carneiro, and Andrew P. Bradley. 2015. *Deep Learning and Structured Prediction for the Segmentation of Mass in Mammograms.* Springer International Publishing, Cham, 605–612. DOI:https://doi.org/10.1007/978-3-319-24553-9_74

[9] N. Dhungel, G. Carneiro, and A. P. Bradley. 2015. Deep structured learning for mass segmentation from mammograms. In *2015 IEEE International Conference on Image Processing (ICIP)*. 2950–2954. DOI:https://doi.org/10.1109/ICIP.2015.7351343

[10] M. G. Ertosun and D. L. Rubin. 2015. Probabilistic visual search for masses within mammography images using deep learning. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 1310–1315. DOI:https://doi.org/10.1109/BIBM.2015.7359868

[11] Krzysztof J Geras, Stacey Wolfson, S Kim, Linda Moy, and Kyunghyun Cho. 2017. High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks. *arXiv preprint arXiv:1703.07047* (2017).

[12] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 1440–1448.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2016. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence* 38, 1 (2016), 142–158.

[14] Richard H Gold, Lawrence W Bassett, and Bobbi E Widoff. 1990. Highlights from the history of mammography. *Radiographics* 10, 6 (1990), 1111–1131.

[15] Michael Heath, Kevin Bowyer, Daniel Kopans, and Richard Moore. The digital database for screening mammography.

[16] Nehmat Houssami, Petra Macaskill, Daniela Bernardi, Francesca Caumo, Marco Pellegrini, Silvia Brunelli, Paola Tuttobene, Paola Bricolo, Carmine Fantò, Marvi Valentini, and others. 2014. Breast screening using 2D-mammography or integrating digital breast tomosynthesis (3D-mammography) for single-reading or double-reading–Evidence to guide future screening strategies. *European Journal of Cancer* 50, 10 (2014), 1799–1807.

[17] M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. Diao, C. Igel, C. M. Vachon, K. Holland, R. R. Winkel, N. Karssemeijer, and M. Lillholm. 2016. Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring. *IEEE Transactions on Medical Imaging* 35, 5 (May 2016), 1322–1331. DOI:https://doi.org/10.1109/TMI.2016.2532122

[18] Oeffinger KC, Fontham EH, Etzioni R, and et al. 2015. Breast cancer screening for women at average risk: 2015 guideline update from the american cancer society. *JAMA* 314, 15 (2015), 1599–1614. DOI:https://doi.org/10.1001/jama.2015.12783 arXiv:/data/journals/jama/934597/jsc150008.pdf

[19] Mieke Kriege, Cecile T.M. Brekelmans, Carla Boetes, Peter E. Besnard, Harmine M. Zonderland, Inge Marie Obdeijn, Radu A. Manoliu, Theo Kok, Hans Peterse, Madeleine M.A. Tilanus-Linthorst, Sara H. Muller, Sybren Meijer, Jan C. Oosterwijk, Louk V.A.M. Beex, Rob A.E.M. Tollenaar, Harry J. de Koning, Emiel J.T. Rutgers, and Jan G.M. Klijn. 2004. Efficacy of MRI and Mammography for Breast-Cancer Screening in Women with a Familial or Genetic Predisposition. *New England Journal of Medicine* 351, 5 (2004), 427–437. DOI:https://doi.org/10.1056/NEJMoa031759 arXiv:http://dx.doi.org/10.1056/NEJMoa031759 PMID: 15282350.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[21] Carol H Lee, D David Dershaw, Daniel Kopans, Phil Evans, Barbara Monsees, Debra Monticciolo, R James Brenner, Lawrence Bassett, Wendie Berg, Stephen Feig, Edward Hendrick, Ellen Mendelson, Carl D'Orsi, Edward Sickles, and Linda Warren Burhenne. 2010. 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. *Journal of the American College of Radiology* 7, 1 (Jan. 2010), 18–27.

[22] Constance D Lehman, Robert D Wellman, Diana SM Buist, Karla Kerlikowske, Anna NA Tosteson, and Diana L Miglioretti. 2015. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine* 175, 11 (2015), 1828–1837.

[23] Daniel Lévy and Arzav Jain. 2016. Breast Mass Classification from Mammograms using Deep Convolutional Neural Networks. *arXiv preprint arXiv:1612.00542* (2016).

[24] John M Lewin, R Edward Hendrick, Carl J DâĂŹOrsi, Pamela K Isaacs, Lawrence J Moss, Andrew Karellas, Gale A Sisney, Christopher C Kuni, and Gary R Cutter. 2001. Comparison of full-field digital mammography with screen-film mammography for cancer detection: Results of 4,945 paired examinations 1. *Radiology* 218, 3 (2001), 873–880.

[25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. SSD: Single shot multibox detector. In *European Conference on Computer Vision*. Springer, 21–37.

[26] Howlader N, Noone AM, Krapcho M, Miller D, Bishop K, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, and Cronin KA (eds). 2017. SEER Cancer Statistics Review, 1975-2014, National Cancer Institute. (2017).

[27] JÃžlia E. E. Oliveira, Mark O. Gueld, Arnaldo de A. AraÃžjo, Bastian Ott, and Thomas M. Deserno. 2008. Toward a standard reference database for computer-aided mammography. (2008). DOI:https://doi.org/10.1117/12.770325

[28] Mei-Sing Ong and Kenneth D Mandl. 2015. National expenditure for false-positive mammograms and breast cancer overdiagnoses estimated at $4 billion a year. *Health affairs* 34, 4 (2015), 576–583.

[29] Kersten Petersen, Mads Nielsen, Pengfei Diao, Nico Karssemeijer, and Martin Lillholm. 2014. *Breast Tissue Segmentation and Mammographic Risk Scoring Using Deep Learning.* Springer International Publishing, Cham, 88–94. DOI:https://doi.org/10.1007/978-3-319-07887-8_13

[30] Etta D. Pisano, Constantine Gatsonis, Edward Hendrick, Martin Yaffe, Janet K. Baum, Suddhasatta Acharyya, Emily F. Conant, Laurie L. Fajardo, Lawrence Bassett, Carl D'Orsi, Roberta Jong, and Murray Rebner. 2005. Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. *New England Journal of Medicine* 353, 17 (2005), 1773–1783. DOI:https://doi.org/10.1056/NEJMoa052911 arXiv:http://dx.doi.org/10.1056/NEJMoa052911 PMID: 16169887.

[31] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. DOI:https://doi.org/10.1007/s11263-015-0816-y

[34] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). http://arxiv.org/abs/1409.1556

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[36] Yew-Ching Teh, Gie-Hooi Tan, Nur Aishah Taib, Kartini Rahmat, Caroline Judy Westerhout, Farhana Fadzli, Mee-Hoong See, Suniza Jamaris, and Cheng-Har Yip. 2015. Opportunistic mammography screening provides effective detection rates in a limited resource healthcare system. *BMC Cancer* 15, 1 (2015), 405. DOI:https://doi.org/10.1186/s12885-015-1419-2

[37] A Van Steen and R Van Tiggelen. 2007. Short history of mammography: a Belgian perspective. *JBR BTR* 90, 3 (2007), 151.

[38] Yu-Dong Zhang, Shui-Hua Wang, Ge Liu, and Jiquan Yang. 2016. Computer-aided diagnosis of abnormal breasts in mammogram images by weighted-type fractional Fourier transform. *Advances in Mechanical Engineering* 8, 2 (2016), 1687814016634243. DOI:https://doi.org/10.1177/1687814016634243 arXiv:http://dx.doi.org/10.1177/1687814016634243