Classification of Whole Mammogram and Tomosynthesis Images Using Deep Convolutional Neural Networks

Xiaofei Zhang, Yi Zhang, Erik Y. Han, Nathan Jacobs, Qiong Han, Xiaoqin Wang, and Jinze Liu

Abstract — Mammography is the most popular technology used for the early detection of breast cancer. Manual classification of mammogram images is a hard task because of the variability of the tumor. It yields a noteworthy number of patients being called back to perform biopsies, ensuring no missing diagnosis. The convolutional neural network (CNN) has succeeded in a lot of image classification challenges during the recent years. In this paper, we proposed an approach of mammogram and tomosynthesis classification based on CNNs. We had acquired more than 3000 mammograms and tomosynthesis data with approval from an institutional review board at the University of Kentucky. Different models of CNNs were built to classify both the 2-D mammograms and 3-D tomosynthesis, and every classifier was assessed with respect to truth-values generated by histology results from the biopsy and two-year negative mammogram follow-up confirmed by expert radiologists. Our outcomes demonstrated that CNN-based models we had built and optimized utilizing transfer learning and data augmentation have good potential for automatic breast cancer detection based on the mammograms and tomosynthesis data.

Index Terms — Mammogram, tomosynthesis, convolutional neural network, classification.

I. INTRODUCTION

A s THE most widely recognized malignancy in women, breast cancer causes approximately 40,000 deaths in the U.S each year [1]. Early identification of breast cancer reduces the death rate significantly [2]. Mammography is the most widely used technology to find breast cancer in early stages before patients exhibit symptoms. By exposing a patient’s breasts to the low volume of X-ray radiation, mammography can identify breast cancer because of the different X-ray absorption rates between normal and abnormal tissues. Tumors can be shown as masses, distortions or micro-calcifications in mammogram images [3]. The thick tissue of patients in the mammogram can be also shown as a mass in mammograms, which may overlap with the tumor and make mammography less sensitive. In 2011, FDA approved breast tomosynthesis as a new mammography technology. By taking multiple X-ray images from different angles and reconstructing them to a video, tomosynthesis provides more help to radiologists on identifying abnormalities because it avoids the overlay of dense tissue and tumor mass [4]. Typically, mammograms and tomosynthesis are procured in two standard orientations: Cranio-caudal (CC) and Medial-lateral-oblique (MLO) views during screening. Fig. 1 shows an example of the CC and MLO views of the mammogram of two breasts and multiple slices of the right CC view of tomosynthesis from the same patient.

Screening mammography is the only imaging modality that has been proven to reduce breast cancer mortality [5]. However, to prevent potential miss diagnosis, mammography is also associated with high recall rates and high false-positive results [6]. With current practice, approximately 10% of all women screened for breast cancer are called back for extra work-ups, yet just 0.5% are determined as breast cancer (that is, 5 cancer detected out of 1,000 women screened, or 5 out of the 100 women called back). The utilization of the newly innovated tomosynthesis in conjunction with mammography was appeared to enhance the precision of breast cancer detection [7]. However, manual characterization by radiologists still

Fig. 1. Illustration of 2D mammogram (A) and 3D tomosynthesis (B).
The Convolutional Neural Network has risen as the most powerful machine learning algorithm for image classification. It is outperforming almost all other conventional classification methods and even human ability [10]. The convolutional process in convolutional neural networks can reduce an image containing millions of pixels to a set of small feature maps by reducing the dimension of input data while retaining the most-important differential features. Using convolutional neural networks to classify mammograms is not entirely new. Most of the research work focused on the classification of small patches of suspicious tumors, referred to as region of interest (ROI) [11]. This is typically carved out of the whole images based on either clinical information or automatic segmentation. Lévy and Jain [12] used convolutional neural networks on small patches of mammograms, achieving a maximum accuracy of 93%. Dhungel et al. [13] built a deep learning based method that automatically segments the area of lesions and then classifies the mammogram. Their best results were 0.74 for the whole image, 0.8 for the whole image plus automatically detected small lesion patches, and 0.91 for the whole image plus manually segmented small patches in terms of auROC [13]. In general, by using small abnormality patches, mammograms classification yields good performance but requires very extensive pre-processing work.

A good classification model for whole mammograms would have various advantages, including (a) saving the work of annotating the tumor patch and reducing its related errors, (b) enhancing the utilization of contextual information close to tumors in mammogram, (c) closely representing the real-world clinical practice, and (d) decreasing the patient call-back rate, thus the number of unnecessary tests conducted, without reducing the detection sensitivity. However, it is significantly more difficult to classify the whole image than the small tumor patches because of the expanded size and feature space. Zhou et al. [14] reported an accuracy of 60.90% on whole mammography classification with convolutional neural networks.

In this work, we developed and assessed various convolutional neural network-based models for whole mammogram classification. We additionally present the first breast cancer classification model using 3D tomosynthesis data, a relatively new technology that is only available to 20% of major hospitals in the US. All data were collected at the Department of Radiology, the University of Kentucky with an institutional review board approval (IRB17-0011-P3K). Techniques including data augmentation [15] and transfer learning [16] are combined with convolutional neural network-based models to optimize the performance of the classifiers.

II. ARCHITECTURE OVERVIEW

In our approach, deep convolutional neural networks were used to classify whole-mammography images based on both the 2D mammograms and 3D tomosynthesis data. The working pipeline consists of three stages: data augmentation, transfer learning, and convolutional neural networks. Ten models with different architecture were built, optimized and compared through cross-validation.

A. Data Augmentation

To achieve good performance, deep neural networks usually require a large number of data. However, biomedical datasets like ours contain a relatively small number of samples due to limited patient volume. By generating new data from the original input data, data augmentation can increase the training data size. Many strategies exist for image data augmentation [10], [17]. In this study, we employed a combination of reflection and rotation. For the 2D mammograms, each original image was flipped horizontally. The original and reflected images were then rotated by each of 90, 180, and 270 degrees. Each original image was thus augmented to eight images. For each tomosynthesis sample, the 3D tomosynthesis image sequences as a whole were either horizontally flipped or not flipped, and then randomly rotated 0, 90, 180 or 270 degrees. Such data augmentation generates relevant training samples because tumors may present themselves in various orientations.

The data augmentation can be performed either before the training or during training. Frontloading the augmentation process reduces the running time of the tests but requires 8 times more disk space to store all images. While this is applicable for 2D mammogram images, for the 3D tomosynthesis data, data augmentation was performed during the training phase to minimize storage usage.

B. Transfer Learning

Transfer learning is a method that can re-use the information obtained by a trained model. In the field of image classification, convolutional neural networks [10] trained in the course of successful projects are sometimes published for use by other researchers. Two popular transfer-learning methods involve (a) fine-tuning the parameters in certain layers of the trained convolutional neural networks, or (b) using the trained convolutional neural networks to calculate the feature maps of new types of data.

Mammography data is different from natural image data due to its limited color distribution and structures. However, it can still leverage the basic image features in terms of edges and shapes that can be soundly detected by well-trained convolutional neural network-based models. This study utilizes AlexNet [10], trained with ImageNet [18]. Considering the fact that mammograms differ dramatically from the images in the ImageNet dataset, the trained AlexNet was used only to obtain the feature maps. Each image in the augmented dataset was resized to 832*832, which resolution was chosen with the goal of retaining tumor pixel information. The ImageNet trained AlexNet was deployed to generate the feature maps for the resized images. AlexNet output the feature maps with the shape of 25*25*256. The feature maps were then used in the training of the following shallow convolutional neural networks.
C. Convolutional Neural Network Architectures

We have built different architectures of convolutional neural networks to classify the 2D mammograms and 3D tomosynthesis images [19]. A general shallow convolutional neural network architecture is shown in Fig. 2. Each convolution layer (Conv layer) includes convolution, batch normalization, leaky ReLU and max pooling. All convolutional neural networks used Max pooling with stride 2. The optimizer used is the Adam optimizer [22]. L2 regularization was introduced in the loss function to prevent overfitting [23]. Dropout was also included to improve the model performance [24]. We also adopted two top-performing convolutional neural network architectures, AlexNet [10] and ResNet50 [25], to classify the 2D whole mammograms. Additionally, we have built several models incorporating transfer learning with feature maps learned from AlexNet. Detailed mathematical description of each step is omitted in this paper as they are well established deep learning techniques.

The complete list of architectures is provided in Table 1. During the training phase, learning rate, dropout and L2 regularization beta were tuned with the range of 0.0001 to 0.1, 0.25 to 1 and 0.00001 to 0.1 respectively. The learning rate decay rate of Adam optimizer was set to 0.985 based on the preliminary results. The batch size was determined by two rules: (1) power of two and (2) largest data size can fit the 8 GB memory.

Imbalanced data represent a common problem in machine-learning projects [26]. If imbalances in the training data are not considered, the resulting model generally performs well on the larger class but poorly on the smaller class. The target dataset for this study was classically imbalanced, with roughly 90% of samples representing negative diagnoses. To reduce the imbalance effect, the mini-batches [20] selected during the training phase were restricted to be balanced. During each training epoch, the training data were randomly split into m folds.

\[
m = \frac{N_{\text{pos}}}{n/2}
\]

Where \(N_{\text{pos}}\) denotes the number of positive samples (smaller class) in the training set, and \(n\) is the batch size. In each iteration, all positive samples (\(n/2\) samples) and \(n/2\) randomly selected negative samples of 1-fold training data were fed to train the convolutional neural network.

For the data input, 2D mammograms and their feature maps were read as three-dimensional tensors with a shape defined as length*width*channels. 3D tomosynthesis data and their feature maps were read as four-dimensional tensors with a shape defined as length*width*depth*channels. Here, depth denotes the number of frames of 3D tomosynthesis data, which may vary across tomosynthesis samples. To obtain a fixed input shape, an equal number of frames were selected for each sample. Selected frames start from frame 0 and with an equal

![Fig. 2. Sample convolutional neural network architecture used in this study. Conv layer denotes the convolution, batch normalization, leaky ReLU and max pooling process. Conv layers are followed by fully connected layers (Fully conn) and output layer.](image)

TABLE I

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Transfer Learning</th>
<th>Input Shape</th>
<th>conv1</th>
<th>conv2</th>
<th>conv3</th>
<th>fc1 neurons</th>
<th>fc2 neurons</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D-A1</td>
<td>No</td>
<td>224<em>224</em>3</td>
<td>6@5*5</td>
<td>16@3*5</td>
<td>--</td>
<td>1024</td>
<td>1024</td>
<td>2</td>
</tr>
<tr>
<td>2D-A2</td>
<td>No</td>
<td>224<em>224</em>3</td>
<td>16@3*3</td>
<td>32@3*3</td>
<td>64@3*3</td>
<td>1024</td>
<td>1024</td>
<td>2</td>
</tr>
<tr>
<td>AlexNet*</td>
<td>No</td>
<td>224<em>224</em>3</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>2</td>
</tr>
<tr>
<td>ResNet50*</td>
<td>No</td>
<td>224<em>224</em>3</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>2</td>
</tr>
<tr>
<td>2D-T1-Alex</td>
<td>Yes</td>
<td>25<em>25</em>256</td>
<td>256@1*1</td>
<td>--</td>
<td>--</td>
<td>1024</td>
<td>--</td>
<td>2</td>
</tr>
<tr>
<td>2D-T2-Alex</td>
<td>Yes</td>
<td>25<em>25</em>256</td>
<td>256@1*1</td>
<td>--</td>
<td>--</td>
<td>1024</td>
<td>1024</td>
<td>2</td>
</tr>
<tr>
<td>2D-T3-Alex</td>
<td>Yes</td>
<td>25<em>25</em>256</td>
<td>256@1*1</td>
<td>--</td>
<td>--</td>
<td>512</td>
<td>512</td>
<td>2</td>
</tr>
<tr>
<td>3D-A1</td>
<td>No</td>
<td>128<em>128</em>16*3</td>
<td>16@3<em>3</em>3</td>
<td>32@3<em>3</em>3</td>
<td>64@3<em>3</em>3</td>
<td>1024</td>
<td>1024</td>
<td>2</td>
</tr>
<tr>
<td>3D-T1-Alex</td>
<td>Yes</td>
<td>25<em>25</em>16*256</td>
<td>32@3<em>3</em>3</td>
<td>--</td>
<td>--</td>
<td>256</td>
<td>256</td>
<td>2</td>
</tr>
<tr>
<td>3D-T2-Alex</td>
<td>Yes</td>
<td>25<em>25</em>16*256</td>
<td>256@1<em>1</em>1</td>
<td>--</td>
<td>--</td>
<td>256</td>
<td>256</td>
<td>2</td>
</tr>
</tbody>
</table>

*Detailed architecture is not listed
Layer not applied
interval in one tomosynthesis sample. In this study, the frame number was set to 16 to fit the hardware limitation.

D. Implementation

The convolutional neural networks were implemented using TensorFlow [27]. All the tests were performed on a machine with two groups of four Nvidia GTX 1080 GPUs, each with 8 GB memory.

E. Performance Evaluation

To evaluate the performance of each prediction model, cross-validation was used. The dataset was randomly partitioned into training and testing datasets. The training set was used to train the model; the results of predictions made on the testing set were used to evaluate the performance of the model. The training-testing ratio used in all validation tests was 4:1.

Receiver operating characteristic curve (ROC) [28] is plotted as the true-positive rate versus the false-positive rate at various thresholds. The area under the ROC curve (auROC) is used to measure the performance of a binary classifier. Tradeoffs can be made based on ROC curves to select the most appropriate classification model. When testing the prediction models in this study, a probability of all test samples in each class was calculated. Using each value in the probability set as the threshold, we can derive true-positive rates (TPRs) and false-positive rates (FPRs). These TPR-FPR data were then used to plot the ROC curve and calculate the auROC score.

III. RESULTS

A. Data

High-quality mammogram data from the University of Kentucky Medical Center were obtained with institutional review board approval (IRB 17-0011-P3K). All mammography images were assessed by experienced radiologists. The dataset includes 3,018 negative and 272 positive mammogram images. Each of the positive images contains at least one biopsy-proven malignant tumor. The negative images do not contain malignant tumors confirmed with at least 2-year negative mammogram follow-up assessed by radiologists, but may have benign masses approved by biopsy or established more than 2-year imaging stability. All exams in the dataset were taken in either CC or MLO view or both. Negative images originated from 793 patients, most of which had 4 images taken: namely, CC and MLO views for both left and right breasts. Positive images originated from 125 patients. Most positive patients have two images collected: CC and MLO views of the breast site with a tumor. For each exam, both the 2D mammogram and 3D tomosynthesis results were obtained. The 2D mammograms were provided in 12-bit DICOM format at 3328*4096 resolution. The 3D tomosynthesis images were provided in 8-bit AVI format with a resolution of 768*1024. Table 2 summarizes the dataset used in this study. For each 3D tomosynthesis AVI file, all frames were processed to a set of 8-bit JPEG images so that the transfer learning can be properly applied. The total number of frames for each 3D tomosynthesis exam varies from 21 to 120.

B. Effect of Data Augmentation

Data augmentation increases the size of the training dataset 8-fold. It significantly improves the performance of almost all convolutional neural network (CNN) architectures tested by roughly 0.1 auROC units. Fig. 3 (A) and (B) depict the training loss status of architecture 2D-T2 with and without data augmentation, and Fig. 3 (C) shows the associated ROC curves. The auROC of the test with data augmentation is 0.73 comparing to 0.62 for the test without data augmentation. The training loss converged more smoothly with data augmentation than without. For this reason, all subsequent tests utilized the data augmentation strategy.

C. 2D Mammogram Classification

We evaluated all convolutional neural network (CNN) architectures on 2D mammography images listed in Table 1. The loss converging status during the training phase of all those architectures are shown in Fig. 4. The optimized parameter combination and results of the best shallow-CNN model, the best classic-CNN model, and the best transfer-learning model for 2D mammograms are summarized in Table 3. While classic-CNN models such as AlexNet do generate competitive results, the best architecture seems to be the one leveraging transfer-learning where feature maps derived from ImageNet-trained AlexNet were used for training. For example, 2D-T2-Alex delivered the best auROC approaching 0.73. The result suggests that utilizing the pre-trained model can be more sensitive in detecting key elements such as edges...
and shapes within a mammogram image as well. However, due to the inherent difference between mammogram image and natural images, further training with these features using even a shallow CNN still delivers better classification accuracy than using AlexNet alone.

**D. 3D Tomosynthesis Classification**

We also evaluated three architectures listed in Table 1 designed for 3D Tomosynthesis images. Cross-validation was used to test one model, 3D-A1, on 3D tomosynthesis data, and two models on 3D tomosynthesis feature maps derived from transfer learning. The loss converging status during the training phase of all 3D tomosynthesis classification architectures are shown in Fig. 5. The optimized parameters and auROC scores for the three models are shown in Table 4. Based on the tests, 3D-T2-Alex exhibited the best performance on 3D tomosynthesis feature maps; similar to 2D mammogram images, transfer learning using ImageNet-trained AlexNet was able to improve the performance of 3D tomosynthesis classification models.

**E. Comparison of Classification Results of 2D Mammogram and 3D Tomosynthesis**

Our current results suggest that the 2D mammogram classification model performs slightly better than the 3D tomosynthesis classification model. However, radiologists generally achieve better classification accuracy on 3D tomosynthesis data. One possible explanation for this phenomenon is that this
study used only a subset of the 3D tomosynthesis frames due to memory limitations and the consistent shape requirement of the input. If the discarded frames contained information for diagnosing cancer that the selected frames lacked, then the frame sampling may have contributed to significant information loss. Another possible reason is that the 2D mammograms have better resolution than the 3D tomosynthesis data used in this study, such that the 2D mammograms may benefit from a higher signal-to-noise ratio [29].

### IV. Discussion

This paper reports our preliminary work on developing and optimizing machine learning models for whole image classification of both 2D mammogram and 3D tomosynthesis images. We evaluated 10 different CNN architectures and conclude that combining both data augmentation and transfer learning methods with a convolutional neural network is the most effective in improving classification performance.

We report the first work that studies both 2D and 3D mammography images for breast cancer. Our current work sheds light on how each type of dataset performs when trained independently. But in practice, 2D and 3D images are complementary to each other, where 2D offers high resolution while 3D offers multiple views. One of our future work is to develop an assembled classifier that integrates the 2D mammogram and 3D tomosynthesis data to achieve optimal performance.

3D tomosynthesis has proven to be much more powerful in manual detecting of tumors in clinical practice than conventional 2D imaging. However, 3D tomosynthesis data is much more challenging to deal with, as it often corresponds to a much bigger feature space, requiring a larger training dataset to obtain better performance and requiring more memory space for training. We believe there is still a great opportunity to improve the performance of 3D tomosynthesis image classification model. We are currently collecting more images while simultaneously obtaining more precise annotation of each slice of 3D tomosynthesis data. Typically only a few frames in 3D tomosynthesis images of a positive exam contain the tumor. Using negative frames within a positive exam may mislead the training of the model. In the meantime, we are also investigating alternative strategies, such as RNN model, that can leverage the sequence information among slices to perform classification.

### References


