2017

MAMMOGRAM AND TOMOSYNTHESIS CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS

Xiaofei Zhang
University of Kentucky, xiaofei.zhang@uky.edu
Digital Object Identifier: https://doi.org/10.13023/ETD.2017.364

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Xiaofei Zhang, Student

Dr. Jinze Liu, Major Professor

Dr. Miroslaw Truszczynski, Director of Graduate Studies

MAMMOGRAM AND TOMOSYNTHESIS CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS

THESIS

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science in the
College of Engineering
at the University of Kentucky

By
Xiaofei Zhang
Lexington, Kentucky
Director: Dr. Jinze Liu, Associate Professor of Computer Science
Lexington, Kentucky
2017

ABSTRACT OF THESIS

MAMMOGRAM AND TOMOSYNTHESIS CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS

Mammography is the most widely used method of screening for breast cancer. Traditional mammography produces two-dimensional X-ray images, while advanced tomosynthesis mammography produces reconstructed three-dimensional images. Due to high variability in tumor size and shape, and the low signal-to-noise ratio inherent to mammography, manual classification yields a significant number of false positives, thereby contributing to an unnecessarily large number of biopsies performed to reduce the risk of misdiagnosis. Achieving high diagnostic accuracy requires expertise acquired over many years of experience as a radiologist.

The convolutional neural network (CNN) is a popular deep-learning construct used in image classification. The convolutional process involves simplifying an image containing millions of pixels to a set of small feature maps, thereby reducing the input dimension while retaining the features that distinguish different classes of images. This technique has achieved significant advancements in large-set image-classification challenges in recent years.

In this study, high-quality original mammograms and tomosynthesis were obtained with approval from an institutional review board. Different classifiers based on convolutional neural networks were built to classify the 2-D mammograms and 3-D tomosynthesis, and each classifier was evaluated based on its performance relative to truth values generated by a board of expert radiologists. The results show that CNNs have great potential for automatic breast cancer detection using mammograms and tomosynthesis.

KEYWORDS: Convolutional neural network, Mammogram, Tomosynthesis, Classification, Deep Learning, Transfer Learning

Xiaofei Zhang

Student's Signature

07/19/2017

Date

MAMMOGRAM AND TOMOSYNTHESIS CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS

By

Xiaofei Zhang

Jinze Liu

Director of Thesis

Miroslaw Truszczynski

Director of Graduate Studies

07/19/2017

Date

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Jinze Liu, for her guidance, encouragement, inspiration, and great support during my thesis study. I would also like to thank my committee members: Dr. Nathan Jacobs, for his help with deep learning throughout the study, and Dr. Xiaoqin Wang, for his assistance with data collection and radiology interpretation. Thanks to Yi Zhang for his help with deep learning. Thanks to Corrine Elliott for her help on thesis correction. The knowledge and experience I gained in working with each of you will be invaluable to my future.

Finally, I would like to thank my wife, Zheng Cui, for her love, encouragement and selfless support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. Introduction

1.1. Breast cancer status

The term “breast cancer” encompasses all forms of cancer that develop from breast tissues, including skin, fibrous tissue, glands, and fat. The breast organ is not essential for human life, for which reason *in situ* breast cancer usually is not fatal. However, cancer cells may fall off of breast tissue and spread to other places in the human body through the blood or lymph fluid system.

In the United States, 99% of breast cancer occurs in women. Roughly 12% of women in the U.S. develop breast cancer during their lifetime. Today, breast cancer is the first place of malignant cancer in women. The death rate of breast cancer has decreased during recent decades. However, because of the large number of patients, approximately 40,000 breast cancer patients die each year in the U.S[1].

When cancer is detected early, the cancer cells are most likely found in an isolated part of the body, making it relatively easy to control with proper treatment. Cancer is much more difficult to cure when the cancer cells have spread to multiple parts of the body. To find breast cancer in early stages, before patients exhibit symptoms, women are recommended to undergo a screening test, commonly a mammogram. The results indicate whether a patient has, or is in high risk of getting, breast cancer. In such cases, further tests (Magnetic Resonance Imaging, Ultrasound imaging, biopsy, etc.) are required to determine the proper treatment[2, 3].

1.2. Mammograms and tomosynthesis

1.2.1. Mammograms

Mammograms are two-dimensional X-ray images of breasts used to detect breast cancer when patients do not exhibit symptoms. Mammography entails exposing a patient's breasts to low levels of X-ray radiation. Breast cancer cells are identifiable from mammograms thanks to the different X-ray absorption rates of normal and abnormal tissues. Tumors appear as masses, while micro-calcification manifests as white dots. However, some breasts have dense tissues that likewise appear in mammograms as masses. In such cases, the tumor mass may overlap with the dense tissue, contributing to false-positive diagnoses. Normally, mammograms capture images in two standard orientations: Craniocaudal (CC) and Medial-lateral-oblique (MLO) during screening. There still a lot of supplementary views used in diagnostic[4].

1.2.2. Tomosynthesis

Breast tomosynthesis is an advanced breast imaging technique first approved by the FDA in 2011. It takes multiple X-ray images at different angles. The images are reconstructed to yield a video from which a radiologist can identify abnormalities. Compared to traditional mammograms, tomosynthesis provides more-accurate results because tumors can be more-easily distinguished from dense tissues using images taken from different angles. In addition, distortion in tissues surrounding the tumor is easier to detect by tomosynthesis, providing proof of tumor malignance[5].

1.2.3 Drawbacks of manual classification of mammograms

Approximately 10% of all women screened for breast cancer are called back for additional work-ups, but only 0.5% are diagnosed with breast cancer (that is, 5 women out of 1,000 screened, or 5 out of the 100 women called back). Thus, ensuring good sensitivity in manual classification of mammograms results in an approximate 95% rate of preliminary false positives. The use of tomosynthesis in conjunction with mammography will improve the accuracy of diagnoses, but manual classification still incurs a high false-positive rate and requires years of experience on the part of the radiologist. These false-positive diagnoses result in an abundance of unnecessary follow-up tests, and thus contribute to increased health-care costs as well as unnecessary emotional turmoil for the patients themselves[2-5].

1.3. Convolutional neural networks

1.3.1. Convolutional neural networks

The convolutional neural network (CNN) is a feed-forward artificial neural network used in pattern classification. Hubel and Wiesel first defined CNNs in the 1960's while studying the neurons of local sensitive and directional selection in the feline cortex. In contrast to the standard artificial neural network, the CNN incorporates a convolving process by which the input of each neuron is connected to the receptive field of the previous layer, and the local feature is extracted. By feeding the training data and adjusting the neuron parameters using back-propagation, a computation to decide the adjustment direction of the parameters in the neurons of each layer, CNN has a unique superiority in speech recognition and image processing with its special structure shared

by local weights.[6, 7] Its layout is closer to the actual biological neural network. Weight sharing reduces the complexity of the network, especially the multi-dimensional. The input vector of the image can be entered directly into the network. This feature avoids the complexity of data reconstruction during feature extraction and classification[6, 8].

1.3.2. Performance of CNNs in image classification

Deep learning with CNNs has emerged as one of the most powerful machine-learning tools in image classification, surpassing the accuracy of almost all other traditional classification methods and even human ability. The convolutional process can simplify an image containing millions of pixels to a set of small feature maps, thereby reducing the dimension of input data while retaining the most-important differential features[6].

The MNIST data set is a large database of handwritten digits containing 60,000 training images and 10,000 testing images. Acting on this database, CNNs have achieved an error rate of only 0.21%, which is close to the human performance of ~0.2% error[9]. Another image dataset, entitled Street View House Numbers (SVHN), contains 600,000 digital images. The best-performing CNN achieved an error rate of 1.69%[10], improving upon the estimated human performance of 2% error[11].

1.3.3. CNNs used in mammogram classification

The use of CNNs to classify mammograms is not entirely new. Daniel Lévy et al. used deep CNNs on small patches of mammograms, achieving a maximum accuracy of 93%[12]. Henry Zhou et al. used CNNs on a dataset of whole mammograms, affording 60.90% accuracy[13]. Neeraj Dhungel et al. built a method that automatically segments the area of a mass and then classifies the mammogram. Their best results were 0.76 in terms of AUROC for automatically segmented small patches, and 0.91 for manually segmented small patches[14]. In general, the classification of mammograms using small abnormality patches affords better performance but requires more pre-processing work.

1.4. Motivation of the study

Mammography is the only screening tool that has been proven to reduce breast cancer mortality. However, it is well known that the rate of call-backs by radiologists for mammogram patients is high, largely due to the low signal-to-noise ratio in mammograms and the wide variability in breast cancer imaging. CNNs exhibit superior performance on many image-classification tasks, and a few research projects have demonstrated that CNNs can perform decently well on the specific task of mammogram classification, albeit usually on partial mammograms.

An effective classification model for whole mammograms would offer multiple benefits, including (a) saving the work of annotating partial mammograms, and (b) reducing the patient call-back rate, and thus the number of unnecessary tests conducted, without harming sensitivity.

The goal of this study is to build a good model for whole-mammogram classification using convolutional neural networks, and thus to reduce the false-positive rate of manual classification. The secondary aim is to explore integration of 3-D tomosynthesis to improve the overall accuracy of breast-cancer detection.

Chapter 2. Data and Methodology

2.1. Study workflow

The workflow for this study is shown in Figure 1:

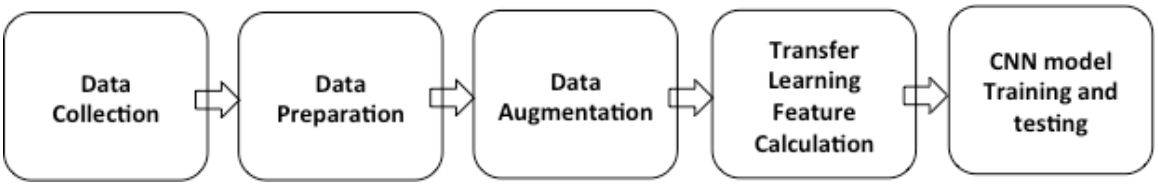

Figure 1: Workflow of this study

2.1.1. Data collection

High-quality mammogram data from the University of Kentucky Medical Center were obtained with institutional review board approval (IRB 17-0011-P3K). The dataset contains 3,018 negative and 272 positive mammogram exams. All positive exams were biopsy-proven malignant cancer samples, and negative exams were assessed by experienced radiologists. All exams in the dataset were taken in either CC or MLO view.

Negative samples originated from 793 patients, for most of whom were collected four images: namely, CC and MLO views for each breast. Positive samples originated from 125 patients. Most positive patients have two images collected: CC and MLO views of the breast site with tumor. For each exam, 2-D mammogram and 3-D tomosynthesis results were obtained. The 2-D mammograms were provided in 12-bit DICOM format at

3328*4096 resolution. The 3-D tomosynthesis images were provided in 8-bit AVI format with a resolution of 768*1024. Table 1 summarizes the dataset used in this study.

Table 1: Mammogram and tomosynthesis data used in this study

View	Negatives	Positives
RCC	758	77
RMLO	759	73
LCC	751	64
LMLO	750	58
Total	**3018**	**272**

2.1.2. Data preparation

All data were de-identified to protect the patients' privacy. In order to save storage space and reduce the time of file I/O, the pixel array for each 2-D mammogram DICOM file was saved as a 16-bit JPEG image. For each 3-D tomosynthesis AVI file, all frames were processed to a set of 8-bit JPEG images for the same purpose. The total number of frames for each 3-D tomosynthesis exam varies from 21 to 120.

2.1.3. Data augmentation

Generally, deep neural networks require training on a large number of training samples to perform well. However, most real-life datasets contain a limited number of samples. Data augmentation is a method for increasing the size of input data by generating new input data based on the original input data. Many strategies exist for data augmentation[6, 15]. This study employed a combination of reflection; rotation by 90, 180, or 270 degrees. For the 2-D 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. Figure 2 depicts the data augmentation process for 2-D mammograms. All augmented images were saved on the hard drive. Compared to executing data augmentation during the training phase, frontloading the augmentation process reduces the running time of the tests. However, for the tomosynthesis data, data augmentation was performed during the training phase due to storage limitations. The data matrix for each tomosynthesis sample was either horizontally flipped or not flipped, and then randomly rotated 0, 90, 180 or 270 degrees.

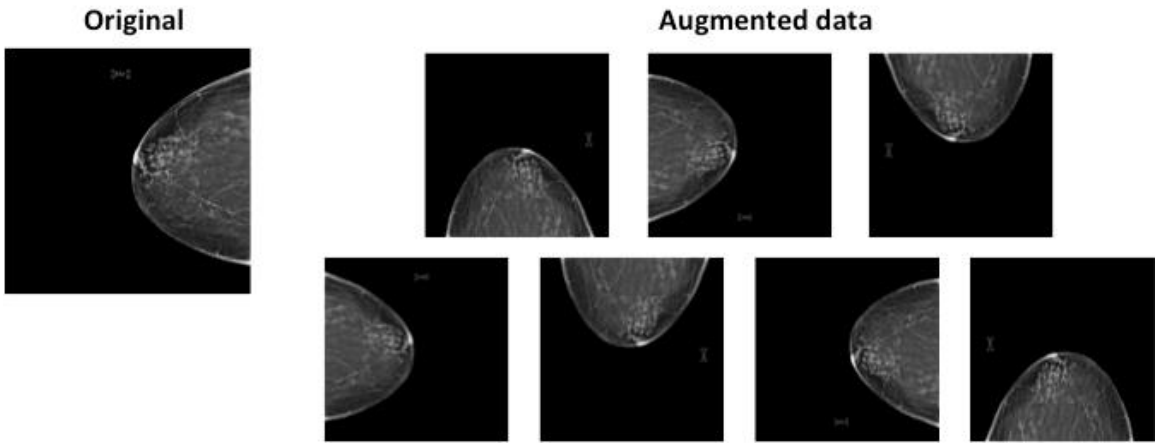

Figure 2: Example of 2-D mammogram data augmentation

2.1.4 Transfer learning

Transfer learning is the re-use of information obtained during the training phase of a previous project. In the field of image classification using CNNs, the CNNs 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 CNN, or (b) using the trained CNN to calculate the feature maps of new types of data.

No mammogram-trained CNN is publicly available, for which reason this study utilizes AlexNet, trained with ImageNet. Figure 3 depicts the architecture of AlexNet[15].

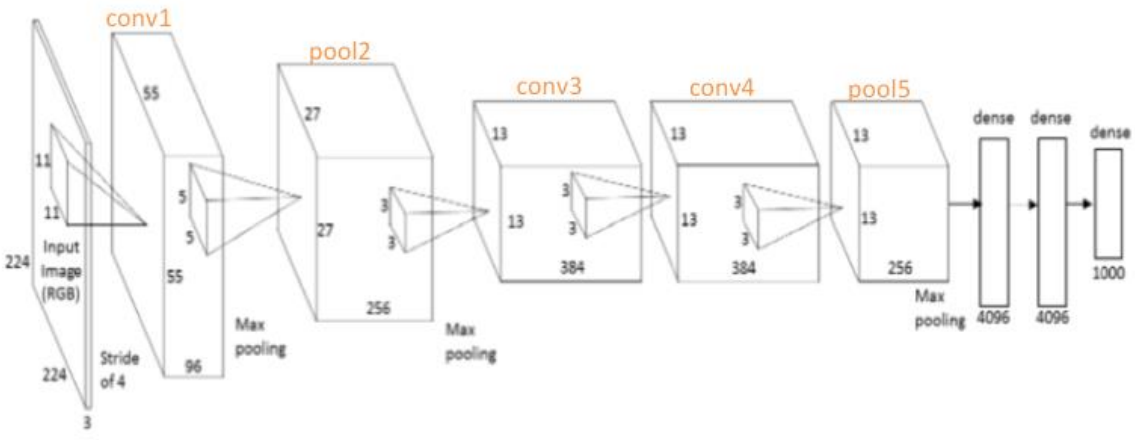

Figure 3: Architecture of AlexNet

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 feature maps (pool5 layer in Figure 3) of resized images were calculated using the ImageNet-trained AlexNet. The output shape is 25*25*256. The feature maps were then used in the training and testing of some shallow CNNs.

2.1.5. CNN models and some details

Different architectures of convolutional neural networks were built to classify the 2-D mammograms, 3-D tomosynthesis images, and their feature maps. Sample CNN architecture is shown in Figure 4. Each convolution process includes convolution, batch

normalization[16], and leaky ReLU[17]. All CNNs used Max pooling with stride 2. The optimizer used is the Adam optimizer[18]. L2 regularization was introduced in the loss function to prevent overfitting[19]. Dropout was also included to improve the model performance[20]. Two classic CNN architectures, AlexNet[15] and ResNet50[21], were also employed to classify the 2-D whole mammograms. Complete architecture details are provided in Table 2.

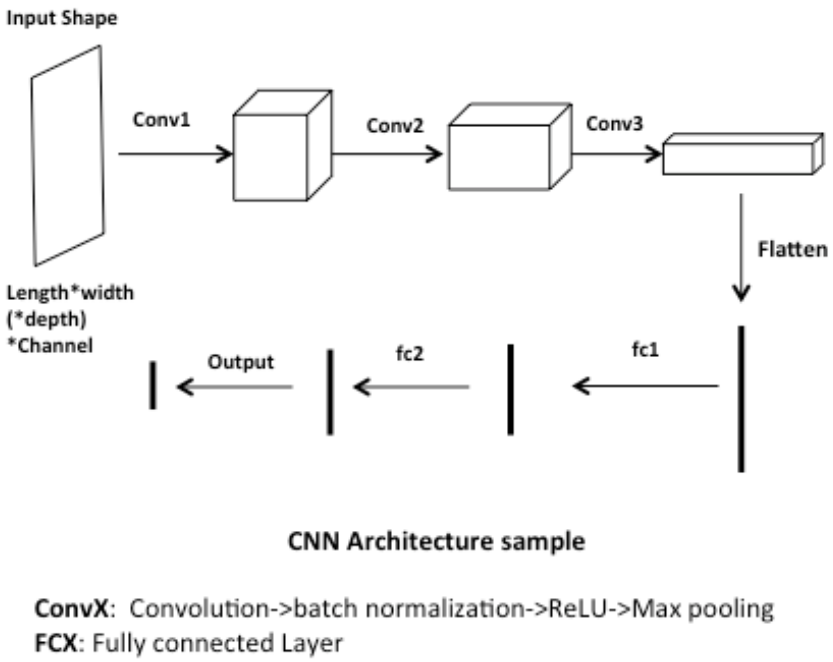

Figure 4: Sample CNN architecture

Table 2: Detailed architectures of tested models for 2-D mammograms and 3-D tomosynthesis classification

Architecture Name	Input Shape	conv1	conv2	conv3	fc1 neurons	fc2 neurons	output
2D-A1	224*224*3	6@5*5	16@3*3	**	1024	1024	2
2D-A2	224*224*3	16@3*3	32@3*3	64@3*3	1024	1024	2
AlexNet*	224*224*3						2
ResNet50*	224*224*3						2
2D-T1	25*25*256	256@1*1	**	**	1024	**	2
2D-T2	25*25*256	256@1*1	**	**	1024	1024	2
2D-T3	25*25*256	256@1*1	**	**	512	512	2
3D-A1	128*128*16*3	16@3*3*3	32@3*3*3	64@3*3*3	1024	1024	2
3D-T1	25*25*16*256	32@3*3*3	**	**	256	256	2
3D-T2	25*25*16*256	256@1*1*1	**	**	256	256	2

*Architectures of Classic CNNs are not shown in table
**The layer was not applied

Among the architectures employed, 2D-A1, 2D-A2, AlexNet, and ResNet50 were used to classify the 2-D whole mammograms. 2D-T1, 2D-T2, and 2D-T3 were used to classify feature maps of the 2-D mammograms calculated by ImageNet-trained AlexNet. 3D-A1 was used to classify whole tomosynthesis samples. 3D-T1 and 3D-T2 were used to classify the 3-D tomosynthesis feature maps calculated by ImageNet-trained AlexNet.

Imbalanced data represent a common problem in machine-learning projects. 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 selected during the training

phase were restricted to be balanced. During each training epoch, the training data were randomly split into *m* folds.

$$m = \frac{N_{pos}}{n/2}$$

Where N_{pos} denotes the number of positive samples (smaller class) in the training set, and *n* is the batch size. On 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 CNN.

For the data input, 2-D mammograms and their feature maps were read as 3-D matrices with shape defined as length*width*channels. 3-D tomosynthesis data and their feature maps were read as 4-D matrices with shape defined as length*width*depth*channels. Here, *depth* denotes the number of frames of 3-D tomosynthesis data, which may vary across tomosynthesis samples. To obtain a fixed input shape, an equal number of frames was selected for each sample. Figure 5 details the method for selecting frames from 3-D tomosynthesis data originally composed of 50 frames.

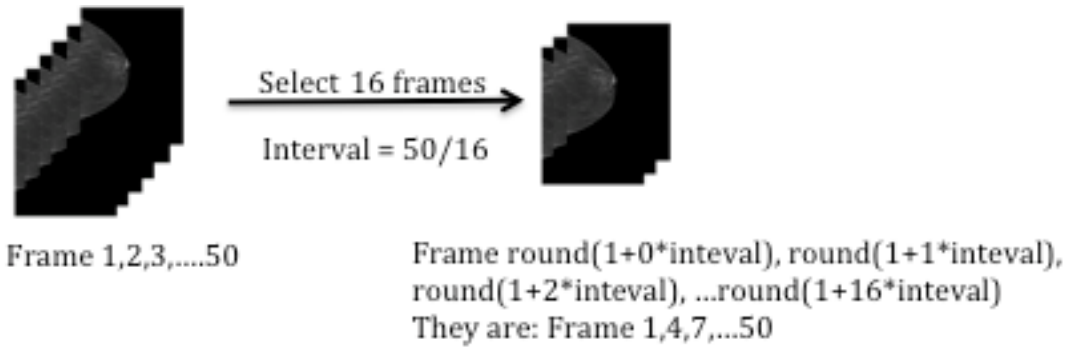

Figure 5: Frame selection method for 3-D tomosynthesis data

2.2. Software and Hardware

2.2.1 Software

The Python programing language and Shell Script were used to build the prediction models and to write testing scripts.

The following libraries were used to build the prediction model:

Deep-learning library:

- Tensorflow 0.10.0[22]
- Tflearn 0.2.1[23]
- Keras 2.0.5[24]

Machine-learning library:

- Scikit-learn 0.18rc2[25]

Image-processing library:

- Pillow 3.4.2[26]
- Pydicom 0.9.9[27]
- Scikit-image 0.12.3[28]

Mathematic library:

- Numpy 1.13.0[29]

All figures were generated using either Gnuplot 5.0[30] or the ggplot2[31] package in R.

2.2.2 Hardware

All deep-learning training and test experiments were run on the University of Kentucky's IPOP Deep Learning server with two groups of four Nvidia GTX 1080 GPUs, each with 8 GB RAM.

2.4. Performance evaluation

2.4.1. Holdout validation and cross-validation

To evaluate the performance of each prediction model, the dataset was partitioned into training and testing datasets. For holdout validation, 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. In this study, the results of holdout validation were used in the parameter-tuning phase and some of the performance-comparison tests. The training-testing ratio used in all holdout validation tests was 4:1.

Cross-validation is a method that provides more-reliable performance evaluation. In this study, 5-fold cross-validation was used to evaluate the performance of models optimized with pre-selected parameters for 2-D mammogram classification. To perform 5-fold cross-validation, the dataset was partitioned into five equally-sized subsets. In each of five tests, four folds were used as training data and the remaining fold was used as testing data. The average result of all five tests was used to gauge the overall performance of the model.

2.4.2. Area Under Receiver Operating Characteristic curve (AUROC)

Receiver operating characteristic curve (ROC) is plotted as the true-positive rate *versus* the false-positive rate at various thresholds. The area under the ROC curve represents the performance of a binary classifier. Tradeoffs can be made based on ROC curves to select the most appropriate model to fit a specific research project. When testing the prediction models in this study, a set of likelihoods of all test samples in each class was calculated. Using each value in the likelihoods set in turn as the threshold, true-positive rates (TPRs) and false-positive rates (FPRs) were calculated. These TPR-FPR data were then used to plot the ROC curve and calculate the AUROC.

Chapter 3. Results and discussion

3.1. 2-D mammogram classification

This section contains the results of 2-D mammogram classification tests.

3.1.1 Parameter-tuning for 2-D mammograms and their feature-map classification models

The augmented 2-D mammogram dataset was used to train and test the CNNs. The results of architecture 2D-A1 using different parameter sets are given in Table 3.

Table 3: Parameter running results of tests of 2D-A1 on 2D mammograms

Tests of 2D-A1	Learning Rate	Dropout	L2 regularization beta	Learning rate decay rate	Holdout AUROC
test1	0.1	0.50	0.001	0.985	0.5488
test2	0.01	0.50	0.001	0.985	0.4737
test3	0.001	0.50	0.001	0.985	0.5026
test4	0.0001	0.50	0.001	0.985	0.4857
test5	0.00001	0.50	0.001	0.985	0.5759
test6	0.1	0.25	0.001	0.985	0.4825
test7	0.1	0.75	0.001	0.985	0.6007
test8	0.1	0.90	0.001	0.985	0.5481
test9	0.1	1.00	0.001	0.985	0.5273
test10	0.1	0.50	0.1	0.985	0.5522
test11	0.1	0.50	0.01	0.985	0.4802
test12	0.1	0.50	0.0001	0.985	0.5513
test13	0.1	0.50	0.00001	0.985	0.5743
test14	0.1	0.50	0.001	0.980	0.5731
test15	0.1	0.50	0.001	0.990	0.5743
test16	0.1	0.50	0.001	0.995	0.5013
test17	0.1	0.50	0.001	1.000	0.5267

These results demonstrate that for the 2-D mammogram dataset, the best parameter set for 2D-A1 is: learning rate = 0.1; dropout = 0.5; L2 regularization beta = 0.001; learning rate decay rate for Adam Optimizer = 0.985.

The results of architecture 2D-A2 using different parameter sets are shown in Table 4.

Table 4: Parameter-tuning results of tests of 2D-A2 on 2-D mammograms

Tests of 2D-A2	Learning Rate	Dropout	L2 regularization beta	Learning rate decay rate	Holdout AUROC
test1	0.1	0.50	0.001	0.985	0.5216
test2	0.01	0.50	0.001	0.985	0.5448
test3	0.001	0.50	0.001	0.985	0.4419
test4	0.0001	0.50	0.001	0.985	0.5429
test5	0.01	0.25	0.001	0.985	0.5382
test6	0.01	0.75	0.001	0.985	0.5051
test7	0.01	0.90	0.001	0.985	0.5015
test8	0.01	1.00	0.001	0.985	0.5488
test9	0.01	1.00	0.1	0.985	0.4782
test10	0.01	1.00	0.01	0.985	0.4923
test11	0.01	1.00	0.0001	0.985	0.5134
test12	0.01	1.00	0.00001	0.985	0.5023
test13	0.01	1.00	0.001	0.980	0.4916
test14	0.01	1.00	0.001	0.990	0.5238
test15	0.01	1.00	0.001	0.995	0.5072
test16	0.01	1.00	0.001	1.000	0.4746

These results suggest that for the 2-D mammogram dataset, the best parameter set for 2D-A2 is: learning rate = 0.1; dropout = 1.0; L2 regularization beta = 0.001; learning rate decay rate for Adam Optimizer = 0.985.

The AlexNet and ResNet50 architectures were tested on the 2-D mammogram data. The results of AlexNet tests are provided in Table 5.

Table 5: Parameter-tuning results of tests of AlexNet on 2-D mammograms

Tests of AlexNet	Learning Rate	Dropout	L2 regularization beta	Learning rate decay rate	Holdout AUROC
test1	0.1	0.50	0.001	0.985	0.5552
test2	0.01	0.50	0.001	0.985	0.4990
test3	0.001	0.50	0.001	0.985	0.5273
test4	0.0001	0.50	0.001	0.985	0.6544
test5	0.00001	0.50	0.001	0.985	0.6256
test6	0.0001	0.25	0.001	0.985	0.6749
test7	0.0001	0.75	0.001	0.985	0.6203
test8	0.0001	0.90	0.001	0.985	0.6279
test9	0.0001	1.00	0.001	0.985	0.6214
test10	0.0001	0.25	0.1	0.985	0.6214
test11	0.0001	0.25	0.01	0.985	0.4864
test12	0.0001	0.25	0.0001	0.985	0.5374
test13	0.0001	0.25	0.0001	0.985	0.6170
test14	0.0001	0.25	0.001	0.970	0.5494
test15	0.0001	0.25	0.001	0.975	0.6274
test16	0.0001	0.25	0.001	0.980	0.6323
test17	0.0001	0.25	0.001	0.990	0.6433
test18	0.0001	0.25	0.001	0.995	0.5634
test19	0.0001	0.25	0.001	1.000	0.5494

These results demonstrate that for the 2-D mammogram dataset, the best parameter set of AlexNet is: learning rate = 0.0001; dropout = 0.25; L2 regularization beta = 0.001; learning rate decay rate for Adam Optimizer = 0.985.

The results for test of ResNet50 are shown in Table 6.

Table 6: Parameter-tuning results of tests of ResNet50 on 2-D mammograms

Tests of ResNet50	Learning Rate	Dropout	L2 regularization beta	Learning rate decay rate	Holdout AUROC
test1	0.1	0.50	0.001	0.985	NA*
test2	0.01	0.50	0.001	0.985	NA
test3	0.001	0.50	0.001	0.985	0.6239
test4	0.0001	0.50	0.001	0.985	0.6111
test5	0.01	0.25	0.001	0.985	0.5948
test6	0.01	0.75	0.001	0.985	0.5267
test7	0.01	0.90	0.001	0.985	0.5387
test8	0.01	1.00	0.001	0.985	0.5575
test9	0.01	0.50	0.1	0.985	0.5262
test10	0.01	0.50	0.01	0.985	0.5731
test11	0.01	0.50	0.0001	0.985	0.5649
test12	0.001	0.50	0.001	0.975	0.5746
test13	0.001	0.50	0.001	0.980	0.4782
test14	0.001	0.50	0.001	0.990	0.6094
test15	0.001	0.50	0.001	1.000	0.5686

*NA: Training loss did not converge

The above results show that for the 2-D mammogram dataset, the best parameter set of ResNet50 is: learning rate = 0.001; dropout = 0.5; L2 regularization beta = 0.001; learning rate decay rate for Adam Optimizer = 0.985.

The original 2-D mammograms dataset was augmented using the method described in Chapter 2, section 2.1.3. The ImageNet-trained AlexNet pool-5 layer's feature maps were calculated using the method described in Chapter 2, section 2.1.4. The shape of the feature map of a single input image is 25*25*256. The test results of architectures 2D-T1, 2D-T2, and 2D-T3 are given in Table 7.

Table 7: Results of tests of transfer learning architectures on 2-D mammogram feature maps calculated by ImageNet trained AlexNet

Architectures	Learning Rate	Dropout	L2 regularization beta	Learning rate decay rate	Holdout AUROC
2D-T1	0.01	0.50	0.0001	0.985	0.6529
2D-T1	0.001	0.50	0.0001	0.985	0.7223
2D-T1	0.0001	0.50	0.0001	0.985	0.6927
2D-T1	0.01	0.50	0.01	0.985	0.7058
2D-T1	0.0001	0.50	0.001	0.985	0.6897
2D-T1	0.0001	0.50	0.00001	0.985	0.7234
2D-T2	0.01	0.50	0.0001	0.985	0.6818
2D-T2	0.001	0.50	0.0001	0.985	0.7274
2D-T2	0.0001	0.50	0.0001	0.985	0.7123
2D-T2	0.01	0.50	0.01	0.985	0.6821
2D-T2	0.0001	0.50	0.001	0.985	0.6691
2D-T2	0.0001	0.50	0.00001	0.985	0.6988
2D-T3	0.01	0.50	0.0001	0.985	0.6737
2D-T3	0.001	0.50	0.0001	0.985	0.7237
2D-T3	0.0001	0.50	0.0001	0.985	0.7085
2D-T3	0.01	0.50	0.01	0.985	0.6967
2D-T3	0.0001	0.50	0.001	0.985	0.7091
2D-T3	0.0001	0.50	0.00001	0.985	0.6827

Based on the results above, the architecture 2D-T2 exhibits the best performance as gauged by holdout validation. The associated AUROC is 0.7274. The best parameter set for 2D-T2 is: learning rate = 0.001; dropout = 0.5; L2 regularization beta = 0.0001; learning rate decay rate for Adam Optimizer = 0.985.

3.1.2 Effect of data augmentation

Three different classification models were tested on the 2-D mammogram feature maps, both with and without data augmentation. The results of these tests are summarized in Table 8.

Table 8: Result of tests using 2-D mammogram feature maps with and without data augmentation

CNN	Augmentation	Best AUROC
2D-T1	No	0.6160
2D-T2	No	0.6162
2D-T3	No	0. 6214
2D-T1	Yes	0.7179
2D-T2	Yes	0.7274
2D-T3	Yes	0.7063

With the help of data augmentation, the performance of each classifier was increased by roughly 0.1 AUROC units. Figure 6 depicts the training loss status of architecture 2D-T2 using 2-D mammogram feature maps. Figure 7 shows the associated ROC curves. The training loss converged more smoothly with data augmentation than without. For this reason, all subsequent tests utilized the data augmentation strategy.

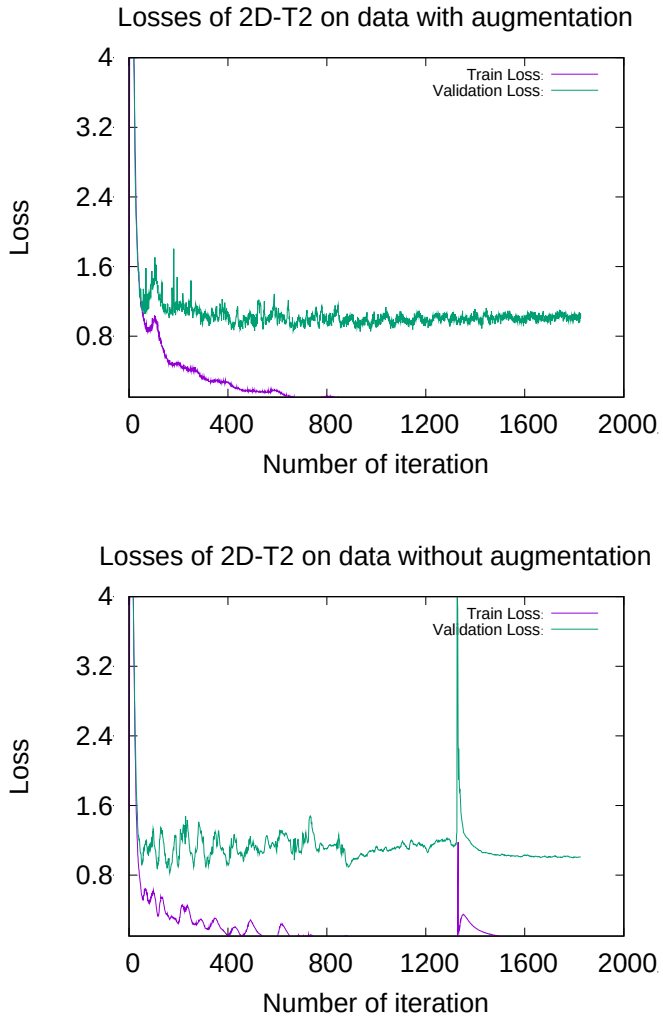

Figure 6: Loss of 2D-T2 on 2-D mammogram feature maps with and without augmentation

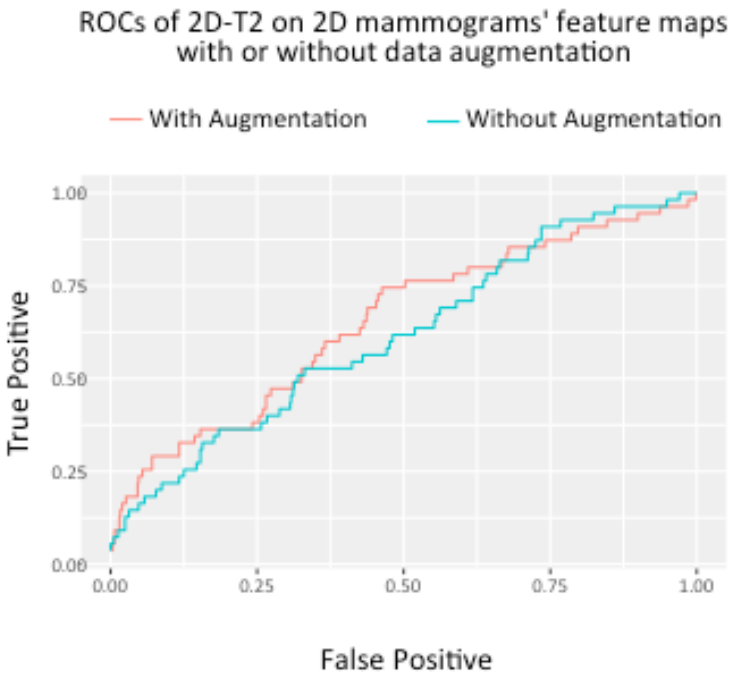

Figure 7: ROC curves of 2D-T2 tested on 2-D mammogram feature maps with and without data augmentation

3.1.3 Summary of 2-D mammogram classification models

The 5-fold cross-validation of the best shallow-CNN model, the best classic-CNN model, and the best transfer-learning model for 2-D mammograms are summarized in Table 9.

Table 9: 5-fold cross-validation results of Shallow CNNs, Classic CNNs and Transfer-learning CNNs

	AUROC					
	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average**
2D-A1	0.5488	0.5480	0.5290	0.5366	0.5431	0.5539
2D-A2	0.5488	0.6030	0.6088	0.6295	0.5495	0.5879
AlexNet	0.6749	0.7048	0.7007	0.6638	0.6294	0.6747
ResNet50	0.6239	0.5299	0.6589	0.5938	0.5995	0.6012
2D-T2	0.7274	0.6522	0. 6741	0.6829	0.6873	0.6848

The best performance was afforded by the transfer-learning architecture 2D-T2 using feature maps calculated by ImageNet-trained AlexNet. The best single-fold AUROC was 0.7274, and the best average AUROC over all five folds was 0.6848. These results suggest that utilizing the transfer-learning strategy can improve the performance of 2-D mammogram classification models.

Figure 8 depicts the training loss convergence status for the best fold tests for each of the five architectures in Table 9.

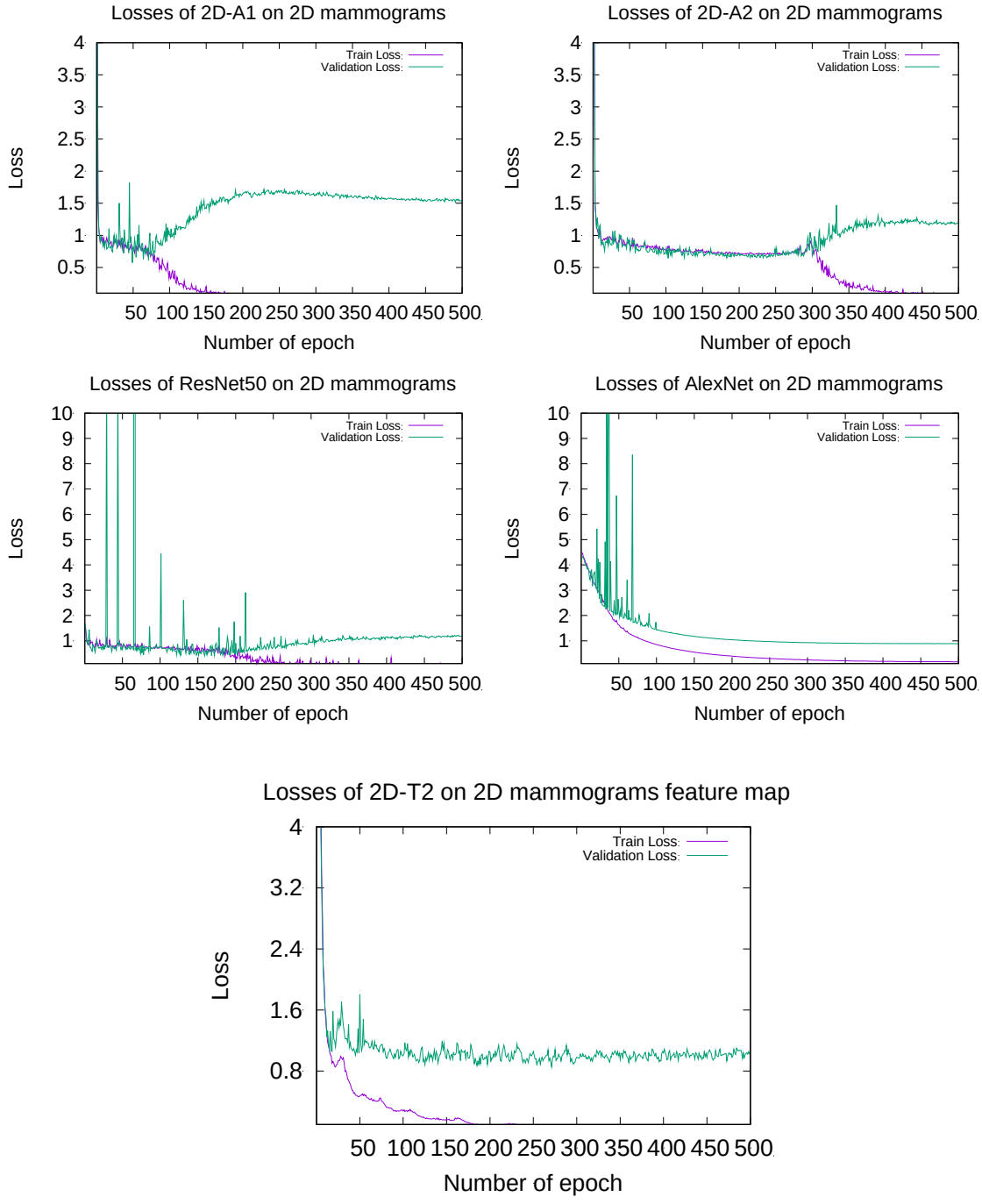

Figure 8: Train loss converge status of five 2-D mammogram classification models

For the architectures 2D-A1, 2D-A2, and ResNet50, the loss status shows that the validation loss increased when the train loss started decreasing. Thus those three models

suffered from overfitting, which explains their low AUROCs. The AlexNet and 2D-T2 architectures afford better loss-converge curves, in which the validation loss did not increase.

3.2. 3-D tomosynthesis classification

3.2.1 Summary of 3-D tomosynthesis classification models

Holdout validation was used to test one model, 3D-A1, on 3-D tomosynthesis data, and two models on 3-D tomosynthesis feature maps. The AUROCs of the best tests for the three models are shown in Table 10, and Figure 9 depicts the associated ROC curves.

Table 10: Holdout validation results of 3-D tomosynthesis classification models

CNN	AUROC
3D_A1	0.6312
3D_T1	0.6116
3D_T2	0.6632

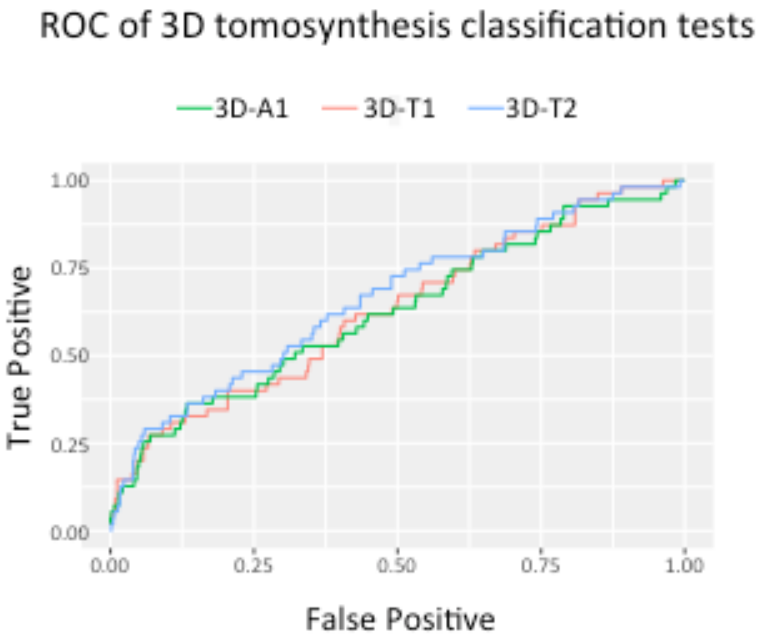

Figure 9: ROC curves of the 3-D tomosynthesis classification models

Based on the tests, 3D-T2 exhibited the best performance on 3-D tomosynthesis feature maps; thus transfer learning using ImageNet-trained AlexNet was able to improve the performance of 3-D tomosynthesis classification models.

Plots of the loss convergence status for tests of the 3-D models are shown in Figure 10. The loss fluctuation at convergence observed for all three models, but especially for the two transfer-learning models, arises due to the small batch sizes used in those tests. Batch size was limited by the memory of the machine used.

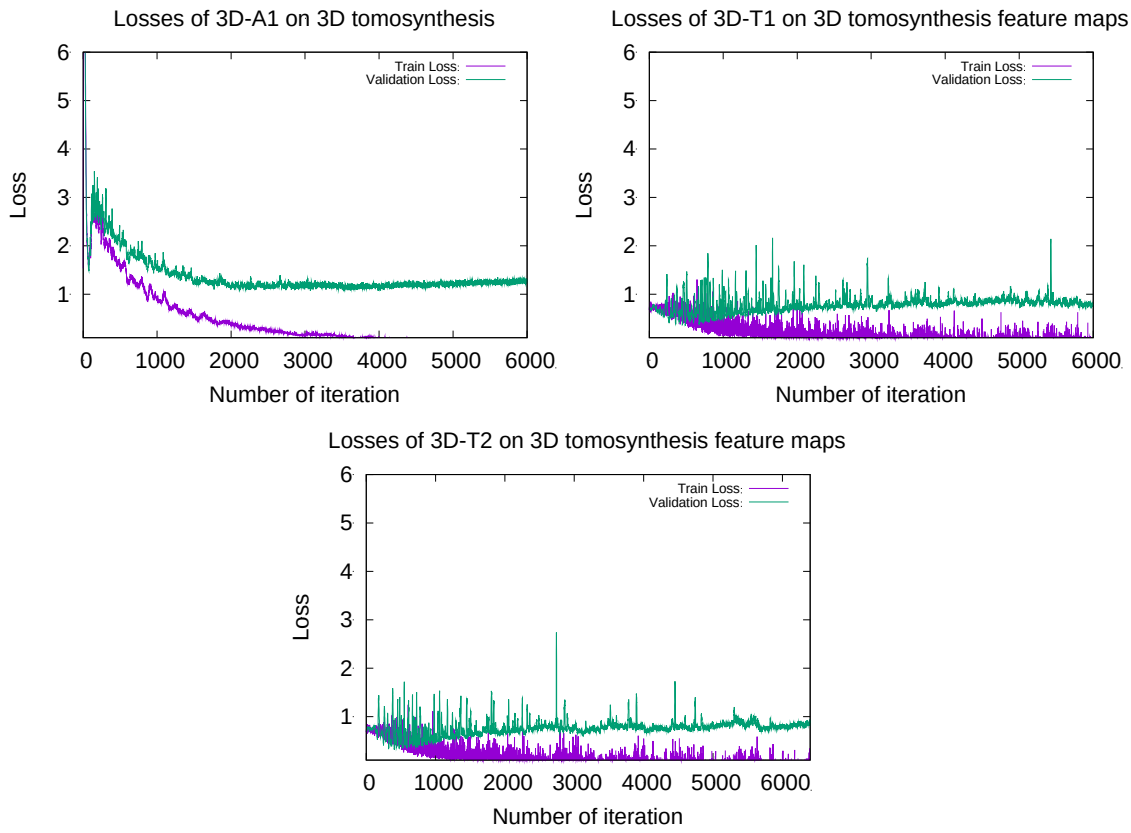

Figure 10: Train loss converge status of three 3-D tomosynthesis classification models

3.3. Comparison of classification results of 2-D mammogram and 3-D tomosynthesis

The best holdout-validation results of 2-D mammogram and 3-D tomosynthesis models were compared. Figure 11 shows their AUROCs, and Figure 12, their loss convergence status.

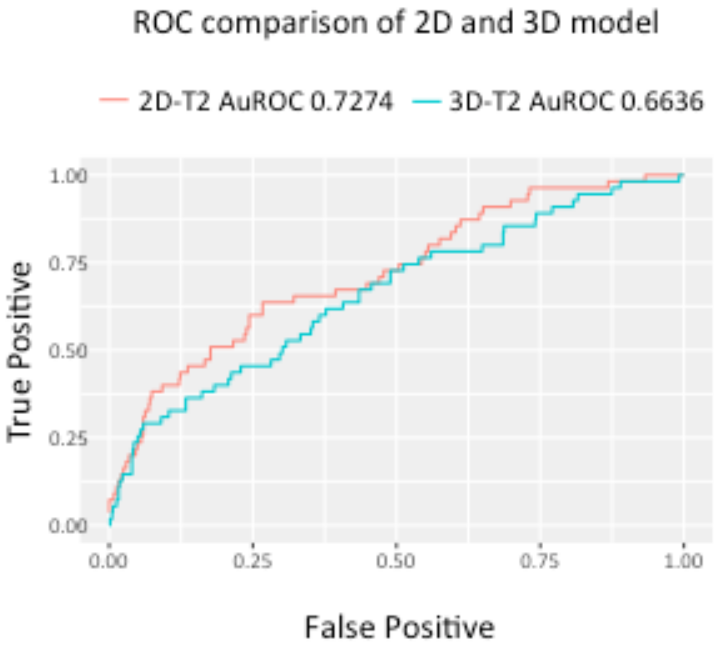

Figure 11: ROCs of best 2-D mammogram classification test and 3-D tomosynthesis classification tests

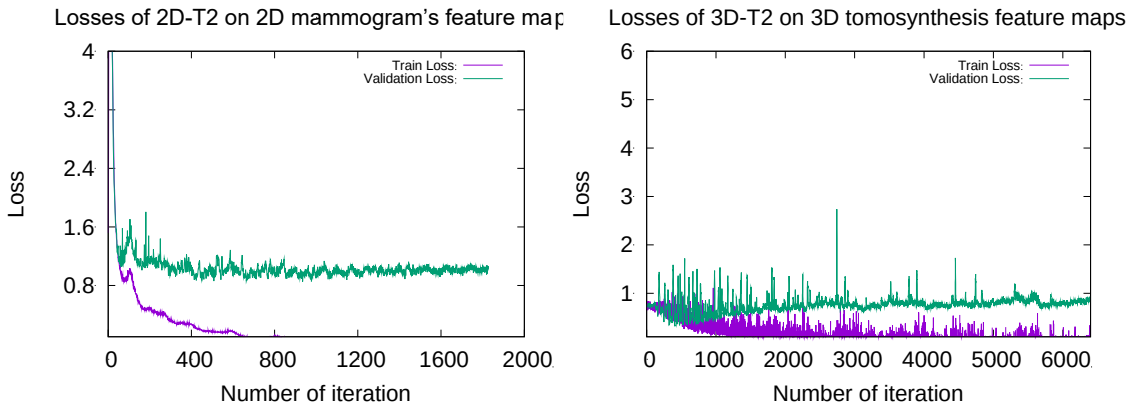

Figure 12: Train loss converge status of best 2-D mammogram classification test and best 3-D tomosynthesis classification tests

These results suggest that the 2-D mammogram classification model performs better than the 3-D tomosynthesis classification model, while radiologists generally achieve better

classification accuracy on 3-D tomosynthesis data. One possible explanation for this phenomenon is that this study used only a subset of the 3-D tomosynthesis frames due to memory limitations. 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 2-D mammograms have better resolution than the 3-D tomosynthesis data used in this study, such that the 2-D mammograms may benefit from a higher signal-to-noise ratio.

Chapter 4. Conclusions and future work

4.1 Conclusions

In this study, we explored classification models using convolutional neural networks on both 2D mammogram and 3D tomosynthesis classification. In the best cases, we achieved AUROC scores of 0.7274 for 2-D mammogram classification, and 0.6632 for 3-D tomosynthesis. The effects of data augmentation and transfer learning were also evaluated, both of which were found to boost the performance of classification models.

4.2 Future work

The current study shows that convolutional neural network models achieved promising results on both 2-D mammogram and 3-D tomosynthesis classification. However, the performance of these models can still be improved.

Firstly, the size of the dataset used in this study is limited. For deep convolutional neural network classifiers, larger datasets usually contribute to better performance. We will first perform tests to determine how many images we should include to improve the performance. Based on the results, more mammography data will be collected.

Secondly, the sampling method of 3-D tomosynthesis frames used in this study may cause information loss. This may be the primary reason that the 3-D tomosynthesis classification model was unable to outperform the 2-D mammogram classification model in this study. A tumor mass may only appear in a few, consecutive frames in the 3-D

tomosynthesis, with the effect that our sampling method might select only one or two frames containing tumor information in 16 frames selected. To address this problem, other sampling methods will be investigated with the aim of reducing information loss during 3-D tomosynthesis data sampling. One thought is to manually annotate the frames containing visible tumors. Another approach is to develop an automatic clustering tool to select out the important frames. Both ideas will be tested in future work.

Finally, more annotation of density levels on the 2-D mammograms can be applied to provide more information for the CNN models. The reason is that tumors in not dense breasts are obvious, while they are not easy to find in dense breasts. With the dense information, the CNN may become more sensitive to distinguish the tumors and dense tissues in mammogram.

Appendix

Abbreviations

AUROC	Area under Receiver Operating Characteristic curve
CNN	Convolutional neural network
CC	Craniocaudal
MLO	Medial-lateral-oblique
ROC	Receiver Operating Characteristic

References

1. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer Statistics, 2017.* CA Cancer J Clin, 2017. **67**(1): p. 7-30.
2. Kim, S.Y., et al., *Breast Cancer Detected at Screening US: Survival Rates and Clinical-Pathologic and Imaging Factors Associated with Recurrence.* Radiology, 2017. **284**(2): p. 354-364.
3. Tosteson, A.N., et al., *Consequences of false-positive screening mammograms.* JAMA internal medicine, 2014. **174**(6): p. 954-961.
4. Poorolajal, J., et al., *Breast cancer screening (BCS) chart: a basic and preliminary model for making screening mammography more productive and efficient.* J Public Health (Oxf), 2017: p. 1-8.
5. Kopans, D.B., *Digital breast tomosynthesis: a better mammogram.* Radiology, 2013. **267**(3): p. 968-9.
6. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-444.
7. LeCun, Y., et al., *Backpropagation applied to handwritten zip code recognition.* Neural computation, 1989. **1**(4): p. 541-551.
8. Bengio, Y., A. Courville, and P. Vincent, *Representation learning: a review and new perspectives.* IEEE Trans Pattern Anal Mach Intell, 2013. **35**(8): p. 1798-828.
9. Wan, L., et al. *Regularization of neural networks using dropconnect*. in *Proceedings of the 30th international conference on machine learning (ICML-13)*. 2013.
10. Lee, C.-Y., P.W. Gallagher, and Z. Tu, *Generalizing pooling functions in convolutional neural networks: Mixed.* Gated, and Tree. arXiv preprint, 2015. **1509**.
11. Netzer, Y., et al. *Reading digits in natural images with unsupervised feature learning*. in *NIPS workshop on deep learning and unsupervised feature learning*. 2011.
12. Lévy, D. and A. Jain, *Breast mass classification from mammograms using deep convolutional neural networks.* arXiv preprint arXiv:1612.00542, 2016.
13. Zhou, H., Y. Zaninovich, and C. Gregory, *Mammogram Classification Using Convolutional Neural Networks.*
14. Dhungel, N., G. Carneiro, and A.P. Bradley. *Deep learning and structured prediction for the segmentation of mass in mammograms*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2015. Springer.
15. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.
16. Ioffe, S. and C. Szegedy. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. in *International Conference on Machine Learning*. 2015.
17. Xu, B., et al., *Empirical evaluation of rectified activations in convolutional network.* arXiv preprint arXiv:1505.00853, 2015.

18. Kingma, D. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.
19. Ng, A.Y. *Feature selection, L 1 vs. L 2 regularization, and rotational invariance*. in *Proceedings of the twenty-first international conference on Machine learning*. 2004. ACM.
20. Srivastava, N., et al., *Dropout: a simple way to prevent neural networks from overfitting.* Journal of Machine Learning Research, 2014. **15**(1): p. 1929-1958.
21. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
22. Abadi, M., et al., *Tensorflow: Large-scale machine learning on heterogeneous distributed systems.* arXiv preprint arXiv:1603.04467, 2016.
23. Tang, Y., *TF. Learn: TensorFlow's High-level Module for Distributed Machine Learning.* arXiv preprint arXiv:1612.04251, 2016.
24. Chollet, F., *Keras*. 2015.
25. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 2011. **12**(Oct): p. 2825-2830.
26. Clark, A., *Pillow (PIL Fork) Documentation*. 2015, Release.
27. Mason, D., *SU-E-T-33: Pydicom: An Open Source DICOM Library.* Medical Physics, 2011. **38**(6): p. 3493-3493.
28. Van der Walt, S., et al., *scikit-image: image processing in Python.* PeerJ, 2014. **2**: p. e453.
29. Walt, S.v.d., S.C. Colbert, and G. Varoquaux, *The NumPy array: a structure for efficient numerical computation.* Computing in Science & Engineering, 2011. **13**(2): p. 22-30.
30. Williams, T., et al., *1 gnuplot.* 2008.
31. Wickham, H., *ggplot2: elegant graphics for data analysis*. 2016: Springer.

Vita

Xiaofei Zhang was born in Yining, Xinjiang, China

Education:

M.S. in Pharmaceutics, University of Peking University Jul. 2012
B.S. in Pharmacy, University of Peking University Jul. 2010

Publications:

1. Xiaofei Zhang, Amir Kucharski, Wibe A. de Jong, Sally R. Ellingson: "Towards a better understanding of on and off target effects of the lymphocyte-specific kinase LCK for the development of novel and safer pharmaceuticals." Procedia Computer Science Volume 108, 2017, Pages 1222-1231

2. Xiaofei Zhang, Sally R. Ellingson: "Computationally Characterizing Genomic Pipelines Using High-confident Call Sets". Procedia Computer Science. Volume 80, 2016, Pages 1023-1032