2020

Multi-Modal Medical Imaging Analysis with Modern Neural Networks

Gongbo Liang
University of Kentucky, tliang130@gmail.com
Author ORCID Identifier:
https://orcid.org/0000-0002-6700-6664

Digital Object Identifier: https://doi.org/10.13023/etd.2020.429

Recommended Citation

Liang, Gongbo, "Multi-Modal Medical Imaging Analysis with Modern Neural Networks" (2020). *Theses and Dissertations--Computer Science*. 100.
https://uknowledge.uky.edu/cs_etds/100

Gongbo Liang, Student

Dr. Nathan Jacobs, Major Professor

Dr. Zongming Fei, Director of Graduate Studies

Multi-Modal Medical Imaging Analysis with Modern Neural Networks

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for the
degree of Doctor of Philosophy in the
College of Engineering at the
University of Kentucky

By
Gongbo Liang
Lexington, Kentucky

Director: Dr. Nathan Jacobs, Associate Professor of Computer Science
Lexington, Kentucky 2020

https://orcid.org/0000-0002-6700-6664

ABSTRACT OF DISSERTATION

Multi-Modal Medical Imaging Analysis with Modern Neural Networks

Medical imaging is an important non-invasive tool for diagnostic and treatment purposes in medical practice. However, interpreting medical images is a time consuming and challenging task. Computer-aided diagnosis (CAD) tools have been used in clinical practice to assist medical practitioners in medical imaging analysis since the 1990s. Most of the current generation of CADs are built on conventional computer vision techniques, such as manually defined feature descriptors. Deep convolutional neural networks (CNNs) provide robust end-to-end methods that can automatically learn feature representations. CNNs are a promising building block of next-generation CADs. However, applying CNNs to medical imaging analysis tasks is challenging. This dissertation addresses three major issues that obstruct utilizing modern deep neural networks on medical image analysis tasks—lack of domain knowledge in architecture design, lack of labeled data in model training, and lack of uncertainty estimation in deep neural networks. We evaluated the proposed methods on six large, clinically-relevant datasets. The result shows that the proposed methods can significantly improve the deep neural network performance on medical imaging analysis tasks.

KEYWORDS: Annotation efficient, weak supervision, network calibration, pre-training, image-text matching

Author's signature: Gongbo Liang

Date: Nov. 16, 2020

Multi-Modal Medical Imaging Analysis with Modern Neural Networks

By
Gongbo Liang

Director of Dissertation: Nathan Jacobs

Director of Graduate Studies: Zongming Fei

Date: Nov. 16, 2020

DEDICATION

This work is dedicated to my parents, my wife, and my daughter. Their inspiration and limitless love and support is the driving force for me to be able to finish my Ph.D. study and curate this document. I would never have become who I am without them.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation and gratitude to my advisor, Dr. Nathan Jacobs, for his constant encouragement, support, and guidance throughout my doctoral program. Dr. Jacobs was always there to listen and give advice when I encountered any issues. He is not only a great mentor who helps me with my research but also an excellent role model who inspires me to work hard. I would have never completed my research without his help, guidance, and support.

I also would like to sincerely thank the rest of my committee members, Dr. Brent Harrison, Dr. Ramakanth Kavuluru, and Dr. Qiang Ye for their valuable feedback and discussions, which helped me curate this document and present my work in a meaningful way. My special thanks go to Dr. Robert Grossman for agreeing to be my outside examiner in my defense. My sincere thanks also go to Dr. Xiaoqin Wang. She provided a handful of discussions and support that helps me conduct this research. I also want to thank Dr. Jin Chen and Dr. Jie Zhang for their help and support during my early Ph.D. studies, which let me enter the world of biomedical imaging research.

I am grateful to the former Director of Graduate Studies, Dr. Mirek Truszczynski, and the Director of Graduate Studies, Dr. Zongming Fei, for their encouragement and support during my Ph.D. studies. I would also like to thank all the faculty members of the computer science department for providing me with a solid foundation, not only in computer science but also in academia.

I am eternally thankful for my former lab-mate, Dr. Zachary Bessinger. As a senior Ph.D. student, Zachary did not only help me in gathering research ideas but also helped me with implementation, coding, and debugging. His help made my research go smoothly, especially when I was a junior Ph.D. student who just started research in deep learning. I am also forever grateful for my other lab-mates, including Tawfiq

Salem, Yu Zhang, Xin Xing, Hunter Blanton, M. Usman Rafique, Connor Greenwell, Weilian Song, and others. I know that I can always ask them for advice, opinions, and ideas. Thank you all for the fun and support. I am eagerly looking forward to having all of you as colleagues in the years ahead.

Last but not least, I would like to thank my family and friends for whose endless love, support, and encouragement have been the light of my life. To my parents, Zhongyi and Zhuoru, when I was a kid, they provided opportunities for me to see varieties of things, places, and cultures in the world, which truly opened my eyes to the world. After I graduated from college, they encouraged me to pursue more education. They are not only my parents who gave me life and raised me but also my abecedarians for my life and career. To my wife, Yang, thank you for the continued support. During the last ten years, she always had faith in me even when I was struggling. Her encouragement and confidence helped me overcome those challenging times. She is still there for me whenever I need her. Without her, I will never complete my Ph.D. and this document. To my daughter, Hannah, she is still very young, but she has brought endless joy to my life. She had already become one of the driving forces for me when she was "in the oven." To my friends, Ray and Marla Flannery, Charlie and Carol Burt, Gene and Ann Perkins, and many other lovely friends, thanks for the endless love, support, and encouragement.

Table of Contents

Acknowledgments . . . . **iii**

Table of Contents . . . . **v**

List of Figures . . . . **vii**

List of Tables . . . . **x**

Chapter 1 Introduction **1**

1.1 Medical Imaging Analysis 2

1.2 Challenges with Convolutional Neural Network 4

1.3 Contributions 5

1.4 Dissertation Outline 5

Chapter 2 Technical Background **7**

2.1 Convolutional Neural Networks 7

2.2 Transfer Learning 9

2.3 Class Activation Mapping 11

2.4 Siamese Networks 12

Chapter 3 Combine Domain Knowledge with Technology **14**

3.1 Introduction 14

3.2 Background 16

3.3 Architecture Overview 18

3.4 Evaluation 21

3.5 Discussion 24

3.6 Conclusion 25

Chapter 4 Learn Fine-Grained Labels from Coarse Labeled Data . . **27**

4.1 Introduction 27

4.2 Architecture . . . 29
4.3 Experiments . . . 33
4.4 Conclusion . . . 34

Chapter 5 Train Deep Learning Models in Low Resource Settings . . . 36
5.1 Introduction . . . 37
5.2 Background . . . 37
5.3 Method . . . 41
5.4 Evaluations and Discussion . . . 43
5.5 Conclusion . . . 52

Chapter 6 A Necessary Step to Safe AI . . . 53
6.1 Introduction . . . 54
6.2 Background . . . 55
6.3 Existing Calibration Methods . . . 56
6.4 Proposed Method . . . 58
6.5 Experiments . . . 59
6.6 Conclusion . . . 66

Chapter 7 Discussion . . . 67

Bibliography . . . 70

Vita . . . 87

LIST OF FIGURES

1.1 An ultrasound image. 1

1.2 Left: breast tumor detection. Red bounding boxes show the detected lesions in the mammogram. Middle: nuclei segmentation. Each circle indicates one segmented nucleus. Right: feature-based CT slices registration. Yellow lines with two circles indicate the matched features in the two CS slices. 3

2.1 A convolutional neural network (CNN) architecture for image classification tasks. There are four types of layers included in this CNN model: convolutional (Conv), pooling (Pool), non-linear activation (rectified linear unit–ReLU), and fully connected (FC) layers. 8

2.2 Illustration of a convolution process of one 3×3 filter (stride 3 and padding 0) from input to output. 9

2.3 A transfer learning model that was pre-trained using the ImageNet data and transferred to a chest x-ray dataset. 10

2.4 Class activation maps (CAM) visualize the decision-making process of a text-imaging matching network on four examples. The CAM heatmap highlights the regions that have a high contribution to the decision of whether the given text and image are matching naturally. 12

2.5 A simple siamese network. 13

3.1 Example of a positive 2D digital mammogram in craniocaudal (CC) view (left) and the four slices from the corresponding digital breast tomosynthesis (right). 15

3.2 Overview of the proposed model. A) DBT data pre-processing, convert DBT data to a fixed representation. B) The input of the model is DM/DBT pairs. C) Feature map extraction using the backbone network. D) Ensemble outputs of different CNN classifiers for testing. Blue lines, model training stage; Green lines, model testing stage; Black lines, shared by the training and testing stage; DBT, digital breast tomosynthesis; DM, digital mammogram. 16

3.3 Example of DM and the corresponding dynamic feature images. 19

3.4 Concatenate features. H=height, W=width, M=modality, C=channel. . 20

3.5 AUC curves of different models. By using 2D digital mammogram (DM) and digital breast tomosynthesis (DBT) data in combination, the model performance can be improved significantly. DM Classifier-Use only DM data for training, DBT Classifier-Use only DBT data for training, DM-DBT Classifier-Use DM and DBT data in combination for training. 24

4.1 Breast tumor localization example of a given input mammogram (left), the prediction of the attention-based method (middle), and the prediction of our method (right). The red box indicates the ground truth tumor localization. The heatmap shows the predictedu location. 29

4.2 An illustration of our weakly-supervised self-training breast cancer localization model: 1) an input image passes through the classification network to extract the intermediate feature maps and CAM heatmap; 2) the localization network is trained using a self-training strategy with the intermediate feature maps and CAM heatmap; 3) at the testing stage, softmax is applied on the fused CAM and localization network outputs to find the final object location. 30

4.3 Breast tumor localization examples generated with our method. The red boxes are the ground truth bounding boxes. The heatmaps show the predicted locations. 35

5.1 An example of a chest x-ray (left) with the radiology report (right) from the MIMIC-CXR dataset. 38

5.2 The TIMNet matching architecture. 1) Weakly-supervised image feature learning through a text-image matching network (solid black line). 2) Downstream application training using a small dataset (dashed blue line). 42

5.3 Examples from different datasets used in evaluation. 44

5.4 Binary classification results on MIMIC-CXR dataset. 46

5.5 Multi-label leanring results on MIMIC-CXR dataset. 47

5.6 Binary classification results on Mendeley-V2 dataset. 48

5.7 Four CAM visualizations for pediatric chest X-ray pneumonia classification on the Mendeley-V2 dataset with chest X-ray on the left and CAM on the right. Top: Two true positive predictions. Bottom: Two false negative predictions. 49

5.8 Multi-class classification results on Kather5000 dataset. 50

5.9 Text and image matching results on the MIMIC-CXR. 50

5.10 Examples of CAM visualizations of text and image matching on MIMIC-CXR. 51

6.1 Left: The train loss/accuracy and test loss/accuracy of the uncalibrated model. The model is overfitted after the 7^{th} epoch, where the train loss keeps decreasing but the test loss keeps increasing. Right: The train loss/accuracy and test loss/accuracy of our method. The DCA term penalizes the model when the loss reduces but the accuracy is plateaued. Both the train and test losses maintain at the same level after the 7^{th} epoch. 54

6.2 The t-SNE plots of the representations learned using temperature scaling and the proposed method on the Mendeley V2 (left) and Kather 5000 (right) datasets. The samples of temperature scaling are spreading in the feature space (left in each subplot). The samples of the proposed method are densely packed for the same class (right in each subplot). 64

6.3 The figure shows the probability distribution that was recovered by the uncalibrated method (left), temperature scaling (middle), and the proposed method (right). The recovered distribution of the uncalibrated model (left) is far from the ground-truth with many overconfident predictions. Temperature scaling (middle) reduces the predicted confidence of the uncalibrated model, but the recovered distribution is still far from the ground-truth. Our method (right) can better recover the true probability. 65

6.4 Train/Test accuracy and loss of Mendeley V2 dataset using AlexNet architecture with four different β values. 66

6.5 Expected calibration error (ECE) of each dataset with different β values. 66

LIST OF TABLES

3.1 Detail of CNN Classifiers. 21

3.2 Ensemble results for different backbone networks. 23

3.3 Comparing with baseline model. 23

3.4 Evaluation results of models trained with a single modality vs. models trained with the multiple modalities. 24

4.1 Localization performances. 34

5.1 Text and image matching results 49

6.1 Datasets used in this study. 60

6.2 Expected Calibration Error (ECE) for Each Model 62

6.3 Maximum Calibration Error (MCE) for Binary Classification Tasks . . . 63

6.4 Maximum Calibration Error for Multi-Class Classification Tasks 63

Chapter 1

Introduction

"Medical imaging encompasses different imaging modalities and processes to image the human body for diagnostic and treatment purposes and therefore plays an important role in initiatives to improve public health for all population groups."

—World Health Organization Administration

The concept of medical imaging began in the year of 1895 when Wilhelm Rontgen, a German professor of physics, invented the X-ray [112]. Over the years, multiple imaging modalities were rapidly developed and adopted [126, 128]. In today's practice, medical imaging serves as a critical device for diagnostic and treatment purposes![73, 93, 127]. It provides visual representations of the interior of a human body and reveals internal structures hidden by the skin and bones [6, 136, 150, 47] (Figure 1.1). Medical imaging is useful to help health practitioners confirm the diagnosis of disease and make decisions regarding treatment plans [36, 132, 31, 90]. However, in many cases, interpreting medical images is time-consuming and challenging even to experienced health providers. Computer-aided diagnosis (CAD) has been brought into clinical practices since the 1980s to help radiologists and physicians on various tasks related to the interpretation of medical images [30, 22, 32]. The majority of the current generation of CADs are built upon traditional computer vision and machine learning techniques [141, 24, 87]. One disadvantage of the current

Figure 1.1: An ultrasound image.

CADs is that they are heavily reliant on hand-crafted features and prior knowledge that may not be transferable among datasets and tasks.

Artificial neural networks (ANN) inspired by the structure and function of human brains introduce a robust end-to-end learning method [77]. In the concept of ANNs, the machine is only given a set of images and labels, and the image features can be learned automatically [102]. As a subset of ANNs, deep convolutional neural networks (CNNs) have shown its power on a wide range of computer vision tasks since the early 2010s. For instance, CNNs surpassed human performance in natural imaging classification tasks in 2016 [45], reported super-human performance on skin cancer detection in 2017 [25], and beat the No.1 world ranked Go player in the same year [123].

Though deep convolutional neural networks have a promising future, many challenges obstruct applying such an advanced technique as one of the core building blocks of the next generation of CADs for medical imaging analysis. In this dissertation, we propose four innovative methods to address three real-world challenges of using CNNs in medical imaging analysis tasks.

1.1 Medical Imaging Analysis

The application of medical imaging analysis may include the following tasks: classification, detection, segmentation, and registration.

Image classification is the task of classifying images according to a set of shared qualities or characteristics. The output of a classification task is a discrete label. A typical classification example will be to decide whether there is a lung nodule in a chest X-ray. The output of this example will be either "yes" or "no." A classification task with two possible answers is also known as binary classification. A classification task with more than two possible answers is known as multiclass or multinomial classification. An example of multiclass classification could be deciding on the subtype of a lung cancer shown in an X-ray. The output of this example will be one of the subtypes.

Object detection is also known as computer-aided detection (CADe) in the medical image analysis field. In addition to classification tasks, the goal of detection is not only to decide whether an abnormality exists in an image but also to find the location of the abnormality in the image. The output of a detection task could be a bounding box that tightly surrounds the abnormality or the x and y coordinates that show the

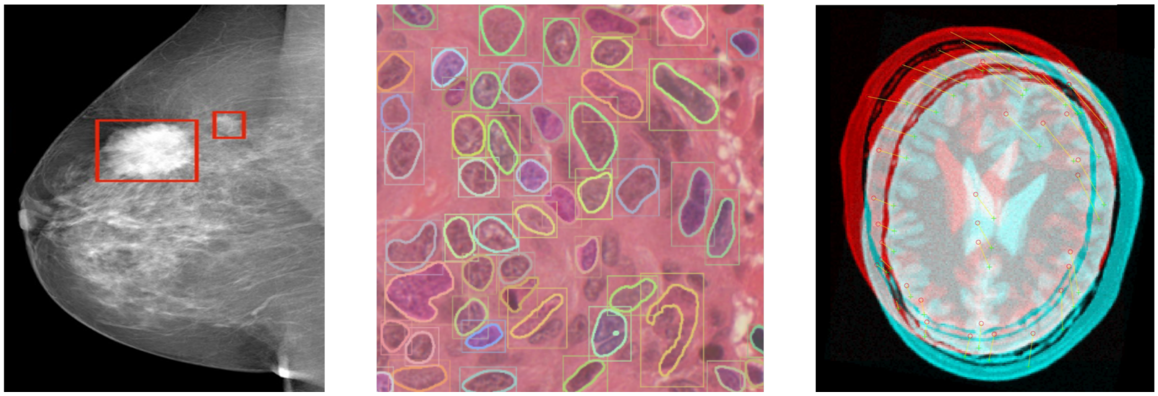


Figure 1.2: Left: breast tumor detection. Red bounding boxes show the detected lesions in the mammogram. Middle: nuclei segmentation. Each circle indicates one segmented nucleus. Right: feature-based CT slices registration. Yellow lines with two circles indicate the matched features in the two CS slices.

abnormality's location. Figure 1.2 (Left) is an example of breast tumor detection. The red bounding boxes show two detected lesions in the mammogram.

The goal of supervised image segmentation is to partition an image into multiple segments. For instance, when given a brain MRI, we may want to strip the skull from the image and leave only the brain. The first step is to partition the pixels of the MRI into one of the two groups, skull or brain. Another example will be that given a pathology image, we want to segment all the nuclei. Figure 1.2 (Middle) is an example of nuclei segmentation. Each circle indicates one segmented nucleus.

Image registration is the process of aligning two or more images into one coordinate system. Images may be captured from different sensors, times, or viewpoints. It is impossible to guarantee that all the images for the same patient appear in the same coordinate system. This misalignment between images could be problematic in automatic medical imaging analysis tasks. Thus, image registration is a preliminary step in many medical imaging analysis tasks that aligns two or more images into the same coordinate system. Figure 1.2 (Right) demonstrates a feature-based registration result of two head CT slices. The yellow line with two circles indicates the matched feature in the two CS slices.

1.2 Challenges with Convolutional Neural Network

A well-known challenge of applying convolutional neural networks (CNNs) on medical imaging analysis tasks is the limitation imposed by the lack of labeled datasets, which is caused by several reasons. Firstly, there is a general shortage of publicly available data. Though millions of medical images are acquired each year, less than a fraction of these images can be publicly accessed due to government regulations and patients' privacy protection. Lack of publicly available data limits the CNN model development and generalization evaluation [85, 152, 147]. Secondly, most of the medical imaging datasets are small, which may not be able to train a CNN model end-to-end without overfitting. For instance, the cancer imaging archive (TCIA), one of the largest collections of cancer images that is accessible for public download, hosts 125 datasets as of October 21, 2020, only five of them contain more than 1000 cases, and a majority have fewer than a hundred cases [7]. Lastly, medical imaging annotation is extremely costly because the annotator needs months or even years of training. Among all the publicly available data, labeled datasets are extremely rare, which obstructs applying supervised methods for CNN training [91, 85, 158].

Another major challenge preventing the use of CNN in clinical practice is the concern of the liability issues, such as who should take the liability of a neural network model or how much health practitioners should trust a model [40, 27, 1]. Researchers are actively developing CNN models that are able to achieve high-performance (e.g., high accuracy); however, uncertainty quantification (i.e., confidence-level of predictions) is often ignored when quantifying these models. Without indicating the confidence of a prediction, the black-boxed result that is provided by a CNN model will be difficult for health practitioners to accept, especially when the modern deep neural networks tend to be overconfident in their predictions [39, 106, 74].

In addition to the challenges mentioned above, knowledge discrepancy between DNN model developers and model users could also be an issue. A model developer is usually someone with strong background knowledge in computer science and may not be as knowledgeable as a health practitioner in handling medical-related tasks. In such a case, less domain knowledge may be applied in the model development process where domain knowledge is extremely critical in many cases.

1.3 Contributions

In this dissertation, we propose four novel deep neural network models to address three real-world challenges of applying CNNs in medical imaging analysis tasks. We also conduct a variety of experiments to show the effectiveness of the proposed methods.

The main contributions of this dissertation are:

- A clinical-inspired framework that integrates domain knowledge into the deep learning model design process. The framework works on both 2D and pseudo-3D medical data simultaneously.
- A weakly-supervised lesion detection method, which uses coarse labels for fine-grained prediction. Through this method, we can relax the annotation requirements of training object detection networks.
- A general framework for image feature learning that utilizes imagery and text data in the medical record system without the request of manual annotations. Various types of downstream applications can be trained using the learned feature representation with only a small labeled dataset.
- A trainable approach to deep neural network classification model calibration, which helps models provide a more accurate estimation of their prediction confidence. The confidence-level of a prediction can be used as a criterion by health practitioners to decide how much to trust a deep learning model.
- A detailed evaluation, both quantitative and qualitative, of the proposed methods on multiple clinically relevant datasets.

1.4 Dissertation Outline

The remainder of this document consists of the following chapters:

- **Chapter** 2 provides a technical background that is necessary for understanding the work in this dissertation. We provide an overview of related background knowledge and research in the convolutional neural net, transfer learning, class activation mapping, and Siamese network.
- **Chapter** 3 introduces a multi-modal classification framework, which evaluates 2D and pseudo-3D data simultaneously. The proposed framework is highly

inspired by the daily clinical practice of specialized breast radiologists. The key technique contribution is that we innovatively convert 3D medical data with a varying number of slices into a fixed-size representation, which significantly reduces the computation of a deep neural network.

- **Chapter** 4 proposes a weakly-supervised lesion detection network. Conventionally, object detection networks are trained fully supervised with bounding boxes or pixel-level annotations. However, such kinds of fine-grained annotations hardly exist in the clinically relevant datasets. In this chapter, we propose a weakly-supervised network for breast lesion detection. The proposed method takes image-level labels as weak supervision and uses a self-learning approach to learn the location of lesions gradually. No bounding boxes or pixel-level labels are needed in this method.

- **Chapter** 5 presents a general framework for image feature learning in a low resource setting. Manual annotations are costly in the medical domain. Deep learning researchers are frequently suffering from the lack of labeled data. Though there is plenty of existing data in the medical record system that provides rich information about medical images, it is still unacceptable to use to directly train a deep learning model. In this chapter, we use unstructured text and image reports as a form of weak supervision. The proposed framework learns image feature representation through an image and text matching network. The learned feature representation can be used to build various downstream applications.

- **Chapter** 6 explores a problem related to safe AI in the medical domain (i.e., how much health practitioners can trust a deep learning model). Modern deep learning networks tend to be overconfident in their predictions, which is problematic in an automatic decision-making system. In this chapter, we propose a trainable approach to deep learning model calibration, which helps models provide a more accurate estimation of the confidence in their prediction.

- **Chapter 7** summarizes the major findings and the contributions of this dissertation. The results of the dissertation will lead to the improvement of medical imaging analysis methods and understanding. Possible future research directions are also discussed.

Chapter 2

Technical Background

In this chapter, to help readers understand the proposed study, relevant technical background information is provided. The concept of convolutional neural networks, a type of neural network commonly used for image-related tasks, is first introduced, followed by the description of transfer learning. Transfer learning is a widely used technique in deep learning model training, especially when the training set is small. Then the method of class activation mapping is discussed, which is a common tool used for deep learning model decision visualization. Finally, this chapter ends with Siamese networks, a neural network that minimizes the intraclass distance and maximizes the interclass distance in the feature space.

2.1 Convolutional Neural Networks

The convolutional neural network (CNN) is a subset of feed-forward artificial neural networks that fall under the concepts of machine learning and artificial intelligence (AI) in a broader sense [117]. CNNs are powerful tools for various imaging-related tasks [72, 44, 25, 133, 165]. For instance, Krizhevsky et al. [72] proposed an eight-layer CNN network that reduces the top-5 classification errors on the ImageNet [20] dataset from 25.8% to 16.4% in 2012. In 2015, He et al. further reduced the figure to 3.57% by assembling multiple very deep CNN models [44], which is higher than human performance (approximately 95%) on this dataset. In the medical field, CNN is widely applied to various imaging analysis tasks, such as medical imaging classification [25], lesion detection [109], imaging segmentation [110], and imaging processing [156].

A CNN model usually contains multiple convolutional (Conv) layers that may be followed by one or more fully connected (FC) layers. The Conv layers learn feature representations of an image, and the FC layers are used for a final decision [130]. Each

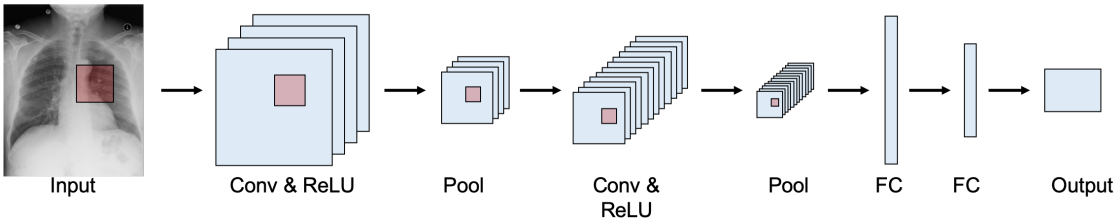


Figure 2.1: A convolutional neural network (CNN) architecture for image classification tasks. There are four types of layers included in this CNN model: convolutional (Conv), pooling (Pool), non-linear activation (rectified linear unit–ReLU), and fully connected (FC) layers.

Conv layer may be followed with a non-linear activation layer and a pooling layer. Figure 2.1 shows an example of a CNN architecture.

A Conv layer contains a set of learnable filters. The size of the filters is usually spatially small, for instance, 3 × 3 or 5 × 5. The filters are moved across the width and height of the input. The convolution process of a Conv layer generates multiple two-dimensional maps (the number of maps is equal to the number of filters). Given a filter (K) and an input image (I), the convolution process with stride 1 can be presented as:

$$conv(I, K)_{x,y} = \sum_{i=1}^{H} \sum_{j=1}^{W} K_{i,j} I_{x+i-1,y+j-1}, \tag{2.1}$$

where H and W are the height and width of I. The output dimension of this filter is:

$$dim(conv(I, K)) = (\left\lfloor \frac{H + 2p - f}{s} + 1 \right\rfloor, \left\lfloor \frac{W + 2p - f}{s} + 1 \right\rfloor), \tag{2.2}$$

where f is the size of the convolutional filter, p is padding, and s is stride. Figure 2.2 illustrates a convolution process of a 3 × 3 filter applied over a 9 × 9 image with stride 3 and padding 0 from input to output. According to Equation 2.2, the output feature map shape is 3 × 3.

The non-linear activation layer applies element-wise non-linearity by using some specific functions. Commonly used non-linearity functions may include:

- Rectified linear unit (ReLU):

$$relu(x) = max(0, x), \tag{2.3}$$

- Hyperbolic tangent function:

$$tanh(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^{2x} - 1}{e^{2x} + 1}, \tag{2.4}$$

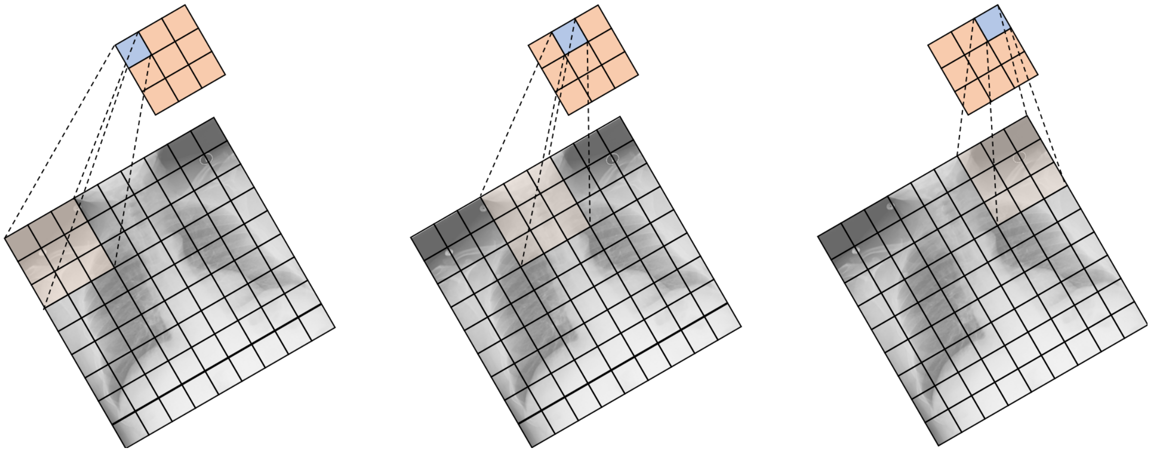


Figure 2.2: Illustration of a convolution process of one 3×3 filter (stride 3 and padding 0) from input to output.

- Sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^x}. \tag{2.5}$$

The pooling layers with stride larger than 1 perform down-sampling, which reduces the spatial size of the representation. The output shape of a pooling layer can be computed using Equation 2.2. A pooling layer is commonly periodically inserted between successive Conv layers. Max-pooling function is one of the most commonly used pooling functions using the maximum value from a region to represent the region. Average pooling is another widely used pooling function.

2.2 Transfer Learning

The recent rapid advancement of the neural networks was enabled by the large amount of data collected during the big data era and advanced computational devices, such as the graphics processing units (GPUs). It is known that the performance of a modern neural network model is associated with the network capacity. For instance, networks with more layers or parameters [124, 137] are more likely to have a better performance than the shallower ones or the ones with fewer parameters [76, 72]. However, the more parameters that need to be learned, the more data that is required during the training stage. Unfortunately, large datasets are rare in many professional domains due to the prohibitively high annotation cost.

Transfer learning is a machine learning technique in which a model trained on one task can be repurposed to improve generalization in another setting [35]. It is an optimization that allows rapid progress or improves performance when modeling

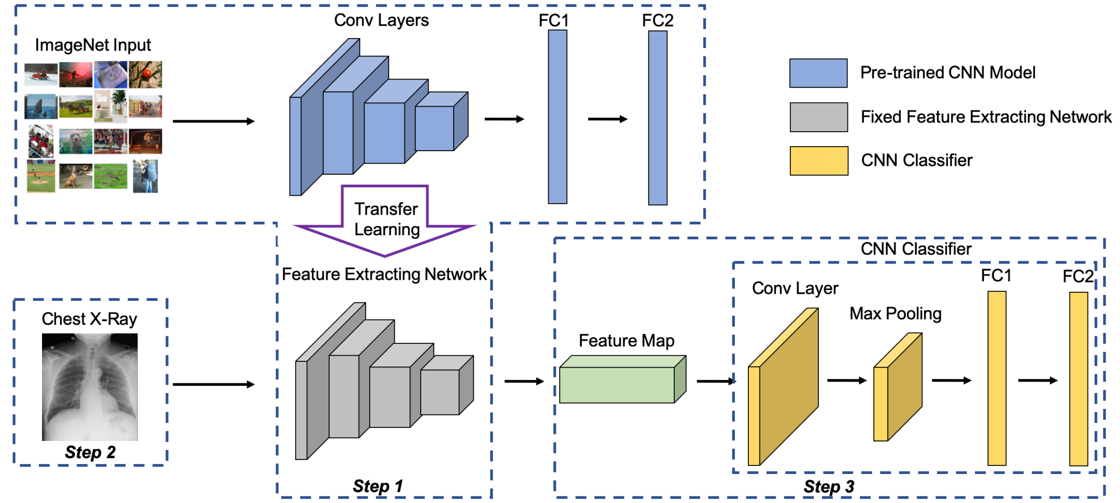


Figure 2.3: A transfer learning model that was pre-trained using the ImageNet data and transferred to a chest x-ray dataset.

the second task [100]. The typical way to do transfer learning in the imaging domain is to pre-train a CNN model on a large dataset. The pre-trained weights are then applied to a second CNN model and train the second model on a new dataset. The pre-trained weights can be applied to all layers, except the last fully connected layer, if the architectures are the same for the two models. The parameters that receive the weights can be either frozen or fine-tuned during the second training stage, but the last fully connected layer must be optimized.

Figure 2.3 illustrates a transfer learning model. The model was pre-trained using the natural imaging dataset and transferred to the second model for a chest x-ray classification task. The parameters of all the convolutional (Conv) layers in the pre-trained model were transferred to the second model. The Conv layers of the second model were frozen and used as a fixed image feature extractor in the second model. A shallow CNN classifier was added after the feature extractor. The shallow CNN classifier contained a new Conv layer, a max-pooling layer, and two fully connected layers. The new Conv layer aimed to convert the natural imaging specific features to chest x-ray specific features. Only the CNN classifier was optimized during the second training stage.

ImageNet dataset [20] (also known as the ILSVRC dataset) is the most used dataset for pre-training. The dataset contains over one million images across 1000 classes. Many researchers report performance improvement when pre-training on ImageNet for numerous tasks in various imaging domains, including the medical imaging domain. However, whether ImageNet is suitable for other imaging domains is still debatable partly because of the significant character differences between the natural imaging domain and others. For instance, a natural image typically has three color channels (R, G, and B channels). However, the majority of the medical images have only one channel. Additionally, objects in a medical image are dramatically different from those in a natural image. Thus, whether we should use ImageNet to pre-train medical imaging analyzing networks still remains an open question.

2.3 Class Activation Mapping

Neural networks are considered as black-boxes. Users have limited information about how a specific decision is made by a neural network. Class activation maps (CAM) is a technique to visualize the decision-making process of a CNN mode [167, 119]. The pixel values of CAM heatmaps are associated with the contribution to the classification decision. A higher value indicates a higher contribution.

Figure 2.4 shows four examples of the decision-making process visualization using CAM heatmaps of a text-imaging matching network. The network is trained to verify whether a given text and image are naturally matching. In all of the examples, the CAM heatmaps successfully highlight the critical parts that associate the image to text. These examples are generated based on [167], in which the CAM heatmap (M_c) for class c can be computed as:

$$M_c(x, y) = \sum_k w_k^c f_k(x, y), \tag{2.6}$$

where (x, y) is a spatial location, $f_k(x, y)$ represents the activation of unit k in the last convolutional layer at the spatial location, w_k^c indicates the weight corresponding to class c for unit k. The bias term is ignored by explicitly setting the input bias of the softmax to 0 since it has little to no impact on the classification performance [167].

More specifically, a classification network with a global average pooling (GAP) layer needs to be trained first. The GAP layer follows the last Conv-layer in the network. After the GAP layer, a fully-connected layer followed by a softmax layer. To generate the CAM heatmap of the predicted class, the following procedures are conducted: 1) retrieve all the weights connected between the fully-connected layer and the softmax class of which we want to predict. If n feature maps are presented before the GAP layer, n weights will be received. 2) Computed the weighted sum of the n feature maps that come from the last Conv-layer. The weighted sum generated a heatmap of a particular class. The size of the heatmap was the same as the feature map.

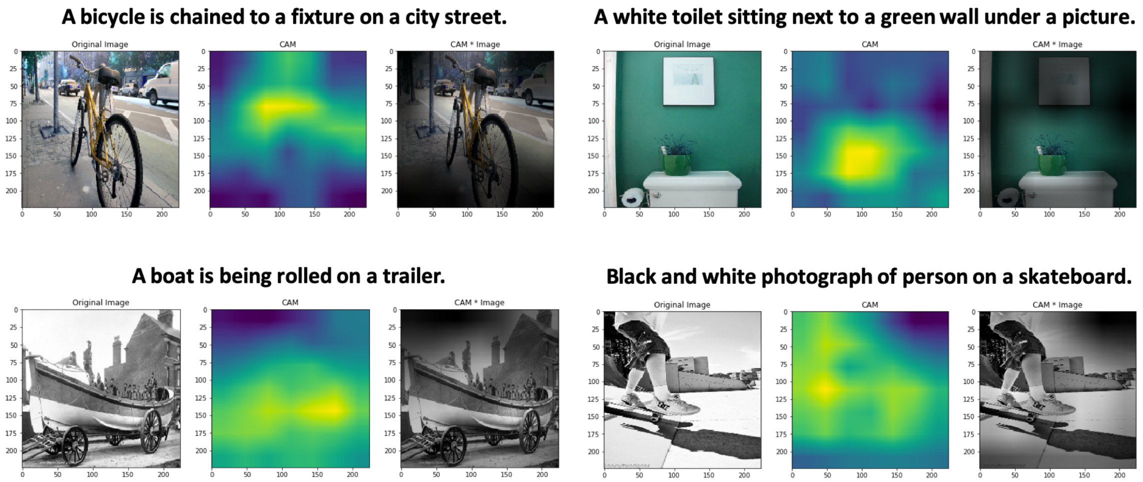

A bicycle is chained to a fixture on a city street.
A white toilet sitting next to a green wall under a picture.
A boat is being rolled on a trailer.
Black and white photograph of person on a skateboard.
Original Image
CAM
CAM * Image
0
25
50
75
100
125
150
175
200

Figure 2.4: Class activation maps (CAM) visualize the decision-making process of a text-imaging matching network on four examples. The CAM heatmap highlights the regions that have a high contribution to the decision of whether the given text and image are matching naturally.

Since the pixel values of CAM heatmaps are associated with the contribution to the classification decision, the pixel values can be interpreted as the possibility of object-of-interest occurrence. Thus, CAMs can also be used as weak supervision for object detection network training [52, 142, 82]. Traditionally, object detection networks are trained with pixel-level labels or bounding box annotations. However, such kinds of fine-grained labels tend to have higher annotation costs than those of image-level labels. By utilizing CAMs, an object detection network using only the image-level labels can be trained.

2.4 Siamese Networks

Siamese neural networks typically contain two identical subnetworks which share the same weights [138, 41]. A siamese network can learn useful data representations that can be used to compare the inputs of the two subnetworks. The outputs of the subnetworks usually are feature vectors. A distance-based loss function is often used to compare the outputs that minimize the loss of samples from the same class and maximize the loss of samples from different classes [18, 118]. The simplest version will be that feeding the absolute difference between the two feature vectors to a fully connected layer for a binary classification task and applying the binary cross-entropy loss.

Figure 2.5 shows a simple siamese network that classifies whether two images are

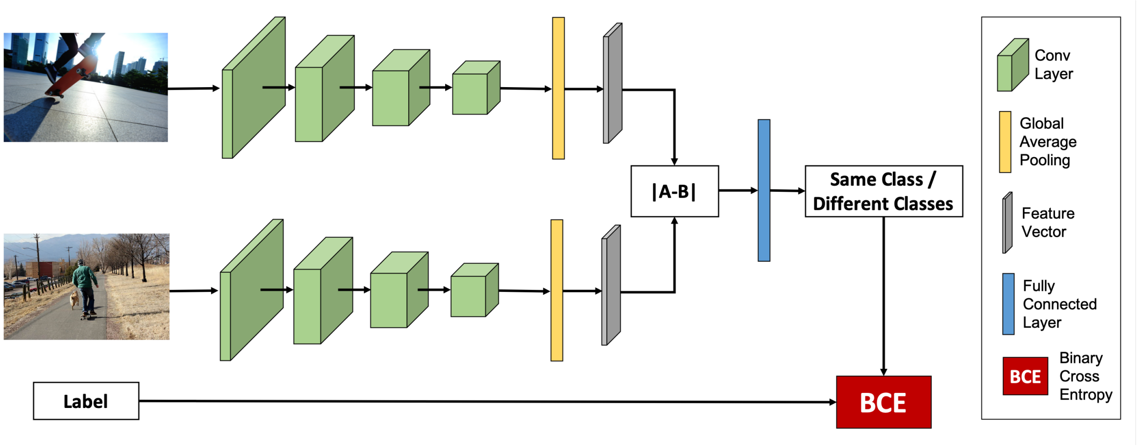


Figure 2.5: A simple siamese network.

from the same class. The network contains two identical image processing subnetworks with shared weights. The image processing network converts an input image to a feature vector. The absolute distance of the two feature vectors are passed to a fully connected layer for classification. Binary cross-entropy loss is used in this example. This loss can be calculated as:

$$Loss_{BCE} = -y\log(p) + (1-y)\log(1-p), \tag{2.7}$$

where y is the class label (i.e., 0 or 1), p is the predicted probability. In order to minimize the loss of samples from the same class and maximize the loss of samples from different classes, both positive and negative samples may be fed at a time and add up the losses of the two samples:

$$Loss = Loss_{BCE_+} + Loss_{BCE_-}, \tag{2.8}$$

where $Loss_{BCE_+}$ is the binary entropy loss for the positive sample, and $Loss_{BCE_-}$ is the binary entropy loss for the negative sample.

Another popular loss function for siamese networks is the triplet loss [118]:

$$Loss_{triplet} = max(d(a,p) - d(a,n) + margin, 0), \tag{2.9}$$

where $d(\cdot)$ is a distance function (such as the $L2$ distance), a is a sample from the dataset, p is a positive sample (e.g., a sample from the same class), n is a negative sample, and $margin$ is an arbitrary margin that helps to distinguish the two pairs of samples better.

A number of applications can be built on siamese networks, such as one-shot learning image recognition [70], face recognition [118], pedestrian tracking [75], resumes-jobs matching [89]. In this dissertation, the idea of siamese networks was used to build a siamese-like network that performs image and text matching tasks.

Chapter 3

Combine Domain Knowledge with Technology

—Clinically inspired deep learning architecture design

In general, most of the development of convolutional neural networks (CNN) in the medical imaging domain is done by computer science experts, who may have a basic understanding of the medical domain but may not be a domain expert. Empirically, domain knowledge is important in medical imaging analysis tools. Thus, there is a need to integrate domain knowledge into CNN model development. This chapter shows that by including domain knowledge from the medical side into CNN architecture design, we can significantly improve the model's performance. We demonstrate this hypothesis through a clinical-inspired multi-modal breast cancer classification network.

3.1 Introduction

Breast cancer is the leading cause of cancer death in over 100 countries and the most frequently diagnosed cancer in 154 out of 195 countries in the world [14, 122]. Mammography is the only image screening tool that has been proven to reduce breast cancer mortality [47]. Digital mammography (DM or 2D mammography) and digital breast tomosynthesis (DBT or 3D mammography) are the two types of mammograms that are used in clinical practice [2] (Figure 3.1). However, the existing models typically focus on using either DM or DBT [109, 162, 92, 166].

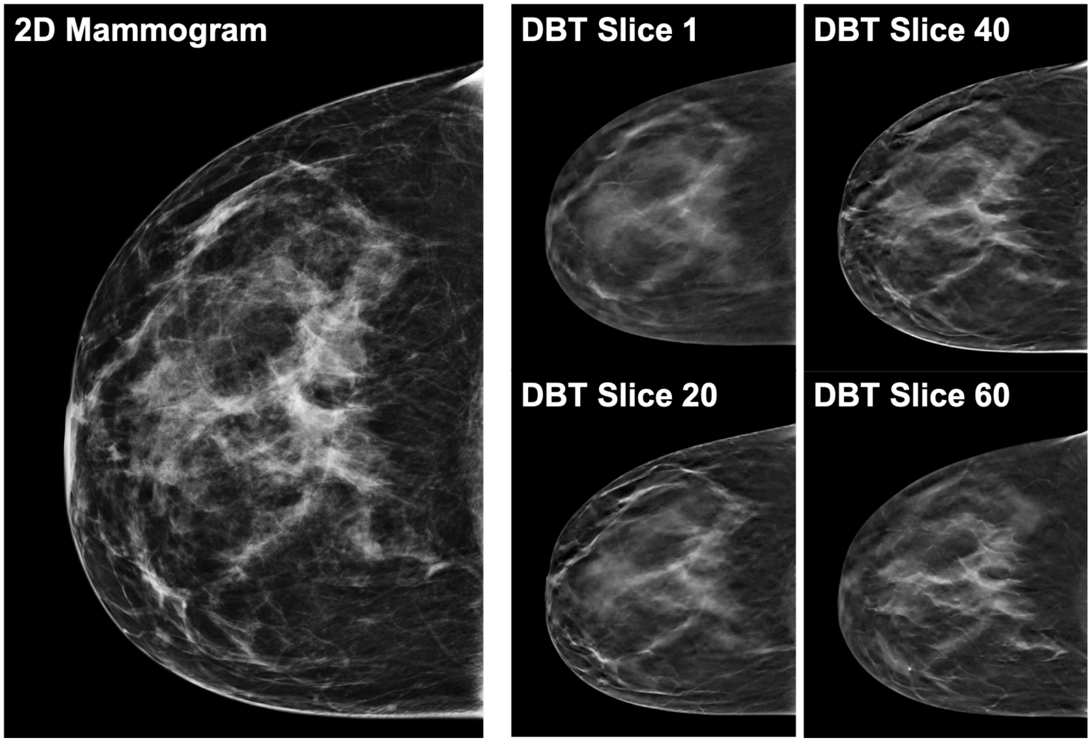


Figure 3.1: Example of a positive 2D digital mammogram in craniocaudal (CC) view (left) and the four slices from the corresponding digital breast tomosynthesis (right).

Inspired by clinical practice, we propose a novel breast cancer classification approach using convolutional neural networks (CNN) that simultaneously reads DM and DBT, as what radiologists would do in their daily practice. One key challenge of this work is how to use DBT effectively. The data size of DBT is large and with varying depths (on average, each DBT has $1024 \times 1024 \times 82$ voxels in this study). Training a 3D CNN model for such large data is extremely computation/memory costly and may potentially lead to overfitting. We innovatively extract a fixed-size slice representation for each DBT and use a 2D CNN for classification. The extracted fixed-size representation can capture the changes between DBT slices, which often helps radiologists evaluate DBT. From our experiment, the proposed method has improved the performance significantly.

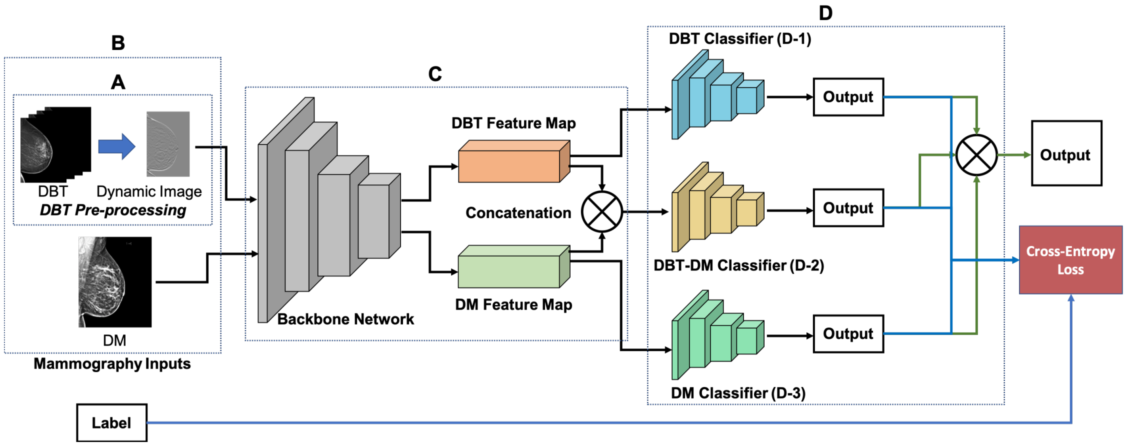


Figure 3.2: Overview of the proposed model. A) DBT data pre-processing, convert DBT data to a fixed representation. B) The input of the model is DM/DBT pairs. C) Feature map extraction using the backbone network. D) Ensemble outputs of different CNN classifiers for testing. Blue lines, model training stage; Green lines, model testing stage; Black lines, shared by the training and testing stage; DBT, digital breast tomosynthesis; DM, digital mammogram.

3.2 Background

3.2.1 Mammography

Mammography is a specialized medical-imaging technique using a low-dose x-ray. It is used in the early detection and diagnosis of breast diseases. Digital mammography (DM), also called full-field digital mammography (FFDM) or 2D mammogram, uses two-dimensional x-ray projections to acquire images with a high spatial resolution (e.g., 3328 × 4096 pixels). While the high resolution helps distinguish the subtle differences between cancer and normal breast tissue, the sensitivity of cancer detection is often limited by superimposed breast tissue, especially in dense breasts [3]. Digital breast tomosynthesis (DBT), also called 3D mammography, is an advanced form of breast imaging, which captures a volumetric view of the breast as a set of parallel 2D image slices, which overcomes the superimposition problem of DM [4, 125]. DBT can lead to better visualization of the underlying tissue; the number of 2D image slices may vary (i.e., from tens to a few hundred) depending on the volume and size of the breast. The large size of DBT data (e.g., on average 1024 × 1024 × 82 voxels in this study) and the varying depth of each DBT can pose a great challenge to computer algorithms. Research has concluded that by considering the slice-to-slice change information of DBT, radiologists' performance can be improved in both accuracy and confidence [94, 105].

3.2.2 CNN-Based Computer-Aided Diagnosis Tools

Ribli et al. used a Faster r-CNN [108] based approach to classify the 2D mammograms, and that achieved 0.95 AUC (area under the receiver operating characteristic curve) for breast tumor classification [109]. This work won second place at the Digital Mammography DREAM Challenge [5]. Shen et al. [121] designed a fully convolutional network for mammogram classification, which achieved 0.94 AUC. Though both the methods reported an exciting performance, the models were trained using bounding boxes (BBs). Such annotations are usually not available on the real clinical data due to the extremely high obtaining cost for medical images. More importantly, these methods are only designed for 2D mammograms. None of them works on DBT.

Mendel et al. [92] proposed a model using a pre-trained VGG19 [124] network as the feature extractor and using a support vector machine (SVM) [19] as the classifier to separately evaluate breast lesions in DM and DBT. They reported 0.81 and 0.89 AUC on DM and DBT, respectively. However, the proposed work has two major limitations. Firstly, a keyframe of each DBT needs to be selected by a trained radiologist during the data pre-processing step. This human involvement does not only increase the cost of using the method but also introduces bias into the proposed model. Secondly, this method omits using the most important informant of DBT—the slice-to-slice changing information, which is the most distinguished feature separating DBT from DM and often used by radiologists for breast cancer detection and diagnosis. Zhang et al. [162] proposed an end-to-end breast cancer classification method using AlexNet [72] as the backbone. Similar to [92], Zhang's method needs to use two different models to evaluate DM and DBT separately. Though this method has some advantages over the previous ones, the model performs poorly on DBT due to the high computational cost of the 3D CNN model. The performance of classifying normal vs. malignant images is only 0.66 AUC. More importantly, neither Mendel's nor Zhang's method is able to evaluate both DM and DBT simultaneously.

3.2.3 CNN Model for Volumetric Data

Two types of 3D CNN models are commonly used to handle volumetric data. The first is the fully 3D CNN architecture, such as 3D-ResNet [43] and I3D [15]. The second is to use 2D CNN models in a 3D way, such as [162]. Even though the two approaches work differently, they both suffer from the same limitations. Due to the CNN model limitation, it is not easy to train a CNN model with a varying size of data. Also, volumetric data usually have much more extensive data size than a regular 2D image, for instance, the average size of ImageNet data is 469×387 pixels, but the average size of DBT used in this study is $1024 \times 1024 \times 82$ voxels. Training a 3D CNN model on such a large data size is extremely computation costly and may potentially lead to overfitting. To reduce the negative effect, Zhang's model only takes 30 slices of each DBT as the input [162]. However, by doing this, either a pre-selection step is needed, or we are just hoping the slices we decide to feed into the model will represent the whole volumetric data sufficiently. Neither of the scenarios is optimal. Thus, directly training a 3D CNN model for DBT may not be a good option.

3.3 Architecture Overview

We propose a novel CNN ensemble method for breast tumor classification. The proposed approach consists of three main components: 1) DBT pre-processing approach (Figure 3.2 A), 2) DBT and DM feature extraction and feature map concatenation (Figure 3.2 C), and 3) multiple classifiers and ensemble outputs of each classifier (Figure 3.2 D).

3.3.1 DBT Pre-processing

In non-medical domains, a popular method to represent a series of images is to apply a temporal pooling operator to the features extracted at individual images. The commonly used temporal pooling operators may include temporal templates [13], ranking functions [29] and sub-videos [49], as well as other traditional pooling operators [114]. We adopt the idea of temporal pooling operators to the medical imaging domain. Inspired by Bilen et al., we applied RankSVM [11] directly on DBT data to extract a fixed, one-slice representation of each DBT. Since the extracted fixed representation keeps the dynamic features (i.e., the slice-to-slice changes) of DBT, we call it *dynamic feature image*. See Figure 3.3 for examples.

One dynamic feature image is a single RGB image, which captures the slice-to-slice changes of a DBT. A ranking function is used to obtain the dynamic feature image for a series of slices I_1,...,I_T, temporally. More specifically, let $\psi(I_t) \in \mathbb{R}^d$ be the feature vector extracted from each individual slices I_t in the series. Let $V_t = \frac{1}{t}\sum_{\tau=1}^{t} \psi(I_\tau)$ be the average time of these features up to time t. The ranking function associates to each time t a score $S(t|\boldsymbol{d}) = \langle \boldsymbol{d}, V_t \rangle$, where $\boldsymbol{d} \in \mathbb{R}^d$ is a vector of parameters. The function parameters $\boldsymbol{d}$ are learned so that the scores reflect the rank of the slices in the series. Therefore, later times are associated with larger scores, i.e. $q \succ t \Rightarrow S(q|\boldsymbol{d}) > S(t|\boldsymbol{d})$.

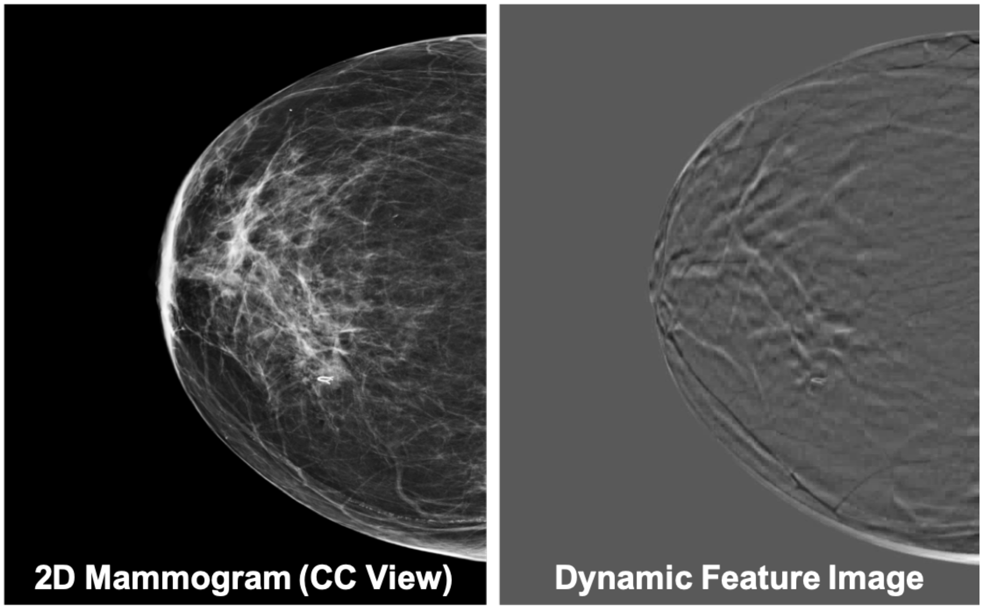


Figure 3.3: Example of DM and the corresponding dynamic feature images.

Learning $\boldsymbol{d}$ is posed as a convex optimization problem using the RankSVM function:

$$\begin{aligned} \boldsymbol{d}^* &= \rho(I_1, ..., I_T; \psi) = \text{argmin}_d\, E(\boldsymbol{d}), \\ E(\boldsymbol{d}) =& \frac{\lambda}{2}||\boldsymbol{d}||^2 + \frac{2}{T(T-1)} \times \\ & \sum_{q>t} \max\{0, 1 - S(q|\boldsymbol{d}) + S(t|\boldsymbol{d})\}. \end{aligned} \tag{3.1}$$

The first term in the objective function is a quadratic regularizer used in SVM. The second term is a hinge-loss that counts how many pairs $q \succ t$ are incorrectly ranked by the scoring function. The optimizer to the RankSVM is written as a function $p(I_1, ..., I_t; \psi)$ that maps a series of T slices to a single vector $\boldsymbol{d^*}$. Since this vector contains enough information to rank all the slices in the series, it aggregates information from all of them and can be used as a descriptor of a series of slices. The process of constructing $\boldsymbol{d^*}$ is known as rank pooling [28]. Rank pooling can be applied directly to the slices of DBT.

3.3.2 CNN Architectures

The proposed network contains two kinds of CNNs: the backbone CNN feature extracting network (Feature Extractor) and the shallow CNN classifier (Classifier).

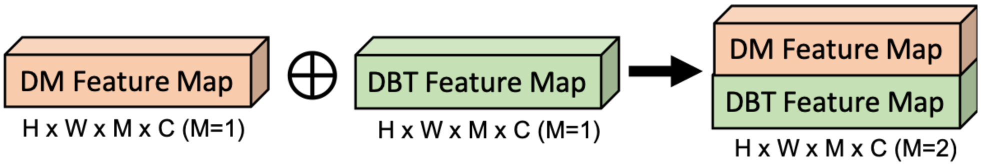


Figure 3.4: Concatenate features. H=height, W=width, M=modality, C=channel.

CNN Feature Extractor

The Feature Extractor is a fully convolutional network (FCN), which takes a $W \times H \times K$ image as input and outputs a $W' \times H' \times K'$ feature map. We use the common CNN classification architecture to build the Feature Extractor by pre-training it on ImageNet [20] dataset. The fully connected (FC) layers of the model are removed. The pooling layer between the first FC layer and the last convolution (Conv) layer is also removed, if applicable. We use the output of the last convolutional layer of the model as the extracted feature map. All the parameters are frozen during the feature extracting step.

CNN Classifier

There are three CNN classifiers with two different architectures included in the proposed model. The DBT Classifier and DM Classifier (Figure 3.2 D-1 and 3.2 D-3) are used for DBT feature map classification and DM feature map classification, respectively. These two classifiers share the same architecture but with different weights, which was implemented as a 2D Conv layer followed by two FC layers. The DM-DBT Classifier (Figure 3.2 D-2) simultaneously evaluates the DM and DBT by taking the feature maps of the two imaging modality in combination and concatenating them on the *modality* dimension (see Figure 3.4). The 2D Conv layer in the other Classifiers is replaced with a 3D Conv layer. The 3D convolution kernels are applied on the *height*, *width*, and *modality* dimensions. All the Conv layers included convolution, batch normalization [54], leaky ReLU [155], and max pooling [72]. The batch size is 32. Max pooling has a 2×2 or $2 \times 2 \times 1$ receptive field with stride 1 for 2D and 3D Conv layers, respectively. Cross-entropy loss is used in training. Adam optimizer [68] with a learning rate of 0.0001 is used as the optimizer. Dropout [131] with a rate of 0.5 is applied to the FC layers. See Table 3.1 for Classifier architecture detail.

Table 3.1: Detail of CNN Classifiers.

Classifier	**Input Shape**	**Conv Layer**	**Conv Type**	**Pooling**	**FC1**	**FC2**
DBT or DM	$w \times h \times c$	c @ 1×1	2D Conv	2×2	256	128
DM-DBT	$w \times h \times 2 \times c$	c @ $1 \times 1 \times 2$	3D Conv	$2 \times 2 \times 1$	256	128

3.3.3 Classifier Ensemble

We propose to use the ensemble learning strategy to improve both the model performance and prediction confidence. In order to keep our method intuitive and straightforward, we use the majority voting strategy [57] in this study.

Suppose we have K classifier, the majority voting can be computed as:

$$C(X) = \arg\max_i \sum_{j=1}^{K} w_j I(h_j(X) = i), \tag{3.2}$$

where h_i is the classifier, w_i is the weights that sum to 1, and $I(\cdot)$ is an indicator function.

3.4 Evaluation

3.4.1 Dataset

A private, clinical dataset is used in this study. All the DM and DBT data were retrospectively collected from patients seen at a comprehensive breast imaging center in the United States from Jan 2014 to Dec 2017. The dataset contains 415 benign patients and 709 malignant patients. Each patient was reviewed by practicing breast radiologists. Both the benign and malignant cases were proved with a biopsy. All patients had both DM and DBT in either craniocaudal (CC) or mediolateral oblique (MLO) view or both views. Approximately 1400 paired DM/DBT data were included. To our best knowledge, this is the largest paired DM/DBT breast cancer dataset.

The DM was provided in 12-bit DICOM format at 3328×4096 resolution. The DBT was provided in 8-bit AVI format with a resolution of 1024×1024. All the frames of every DBT data was saved to a set of 8-bit JPEG images before generating the dynamic feature images. Both the DM and dynamic feature images were down sampled to 832×832. Data augmentation was also applied to each of the mammography images and dynamic feature images through a combination of reflection and rotation. Each original image was flipped horizontally and rotated by each of 90, 180, and 270 degrees. In total, 6875 paired DM/DBT data were used in this study.

The dataset was randomly partitioned into training and testing datasets with a $4:1$ ratio on the patient-level. All the images of the same patient will be in either the training set or the test set. The benign and malignant ratio was maintained in both training and testing sets. To minimize the imbalance effect (low benign to malignant ratio), we balanced each mini-batch during training.

3.4.2 Implementation and Evaluation Metrics

Four popular CNN networks were used as the backbone feature extractor in this study, namely AlexNet [72], ResNet [44], DenseNet [51], and SqueezeNet [53]. The model was deployed in Pytorch [103] and trained with balanced mini-batches for 100 epochs on a Linux computer server with eight Nvidia GTX 1080 GPU cards. The whole training process, including the feature extraction and the CNN classifier training, can be done within several hours.

The classification accuracy (ACC), the area under the receiver operating characteristic curve (AUC), precision (Prec), recall (Reca), F1, average precision (AP), and average correct predict confidence (AC) were used in this study.

3.4.3 Baseline Model and Ensemble Approach

We use the 2D-T3-Alex and 3D-T2-Alex models from [162] as the baseline model for DM and DBT, respectively. The 2D-T3-Alex model is a transfer learning 2D CNN model, which uses pre-trained AlexNet to extract features. The 3D-T2-Alex model is a 3D CNN model, which firstly uses the regular AlexNet model to extract feature maps of every slice in a DBT. Then, K feature maps of each DBT are fed into a one-Conv-layer 3D CNN model for classification. $K = 30$ was chosen in their paper [162].

Our experiment shows the proposed model significantly improves the performance. By only using DBT data (i.e., the dynamic feature images), the performance can be improved from 0.72 AUC to 0.89 AUC (23.61% increase). When using DM and DBT in combination, a single model can achieve 0.95 AUC. After assembling the three classifiers (DM Classifier, DBT Classifier, and DM-DBT Classifier, which uses DM only, DBT only, and DM and DBT data, respectively), the proposed model can further improve the performance to 0.97 AUC (Table 3.2).

Table 3.3 lists the ensemble result of all the different backbone networks. The performance is consistent among the four different feature extractors, which indicates the proposed method is not limited to any specific architecture.

Table 3.2: Ensemble results for different backbone networks.

Model	Input Data	Backbone Network	AUC
$2D-T3-Alex$	DM only	AlexNet	0.87
$3D-T2-Alex$	DBT only	AlexNet	0.72
$Ours_{DBT}$	DBT only	AlexNet	0.89
$Ours_{DM-DBT}$	DM & DBT	AlexNet	0.95
$Ours_{Ensemble}$	DM & DBT	AlexNet	**0.97**

Table 3.3: Comparing with baseline model.

Backbone Network	Input Data	AUC
AlexNet	DM & DBT	0.97
ResNet	DM & DBT	0.96
DenseNet	DM & DBT	0.97
SqueezeNet	DM & DBT	0.97

3.4.4 Single Modality vs. Multiple Modalities

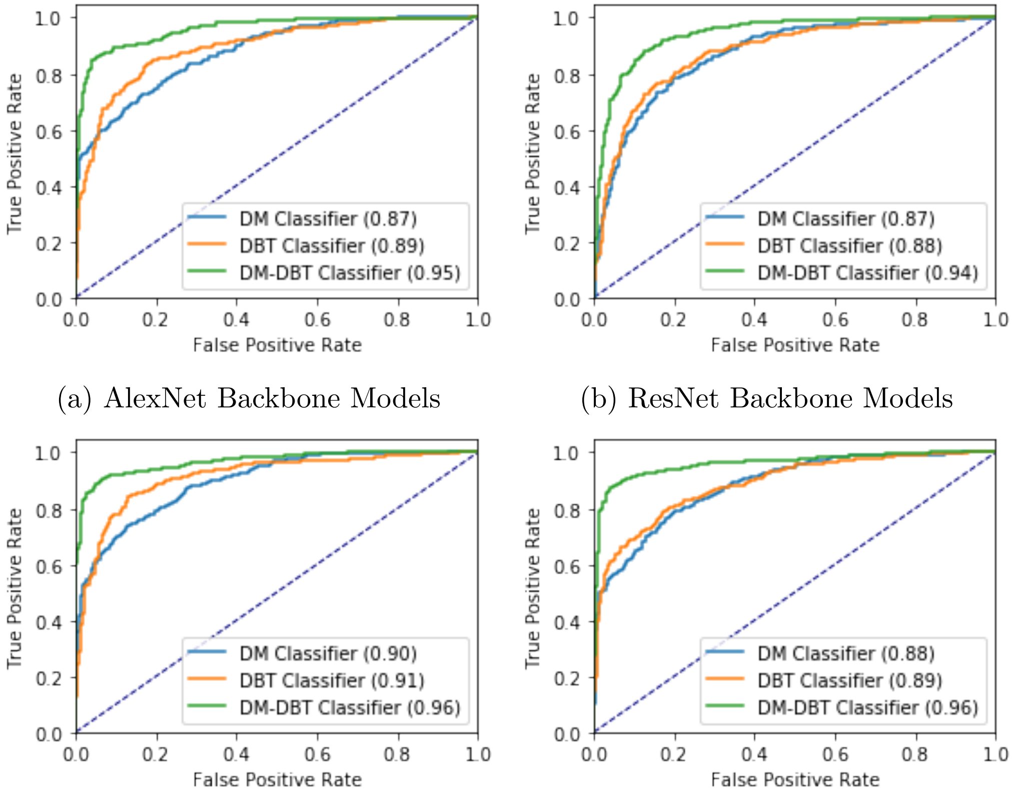
In this section, we evaluate the model performance using a single imaging modality vs. multiple imaging modalities. More specifically, we are comparing the performance of DM Classifier, DBT Classifier, and DM-DBT Classifier. Four different backbone networks were used. In total, 12 models were trained and compared in this experiment.

Table 3.4 reveals when using multiple imaging modalities together, the model performance is significantly better. The DM-DBT Classifier achieves a 0.95 AUC on average. However, the save metric for DM Classifier and DBT Classifier is 0.88 and 0.89, respectively. The table also shows when using DBT data, the model prediction confidence can be improved, especially when using DM and DBT in combination. On average, the prediction confidence of DM Classifier is 0.83, the same metric of DBT Classifier and DM-DBT Classifier is 0.89 and 0.93, respectively. As in the previous section, the performance of all four different backbone networks is consistent. They all achieved a similar result, except the average prediction confidence of single modality classifiers (i.e., DM Classifier and DBT Classifier). Among the four backbone networks, the DenseNet performance is slightly better than others, which achieves the highest scores of 17 out of 21 different metrics for different classifiers.

Figure 3.5 shows the AUC curve of the different models. The green line indicates the AUC performance of DM-DBT Classifier, the orange line is for the DBT Classifier, the blue line is for the DM Classifier, and the dashed line indicates a random performance (0.5 AUC).

Table 3.4: Evaluation results of models trained with a single modality vs. models trained with the multiple modalities.

Backbone	**DM Classifier**							**DBT Classifier**							**DM-DBT Classifier**						
Network	**ACC**	**AUC**	**F1**	**Prec**	**Reca**	**AP**	**AC**	**ACC**	**AUC**	**F1**	**Prec**	**Reca**	**AP**	**AC**	**ACC**	**AUC**	**F1**	**Prec**	**Reca**	**AP**	**AC**
AlexNet	0.78	0.87	0.76	**0.87**	0.75	0.70	0.78	0.81	0.89	0.80	0.84	0.76	0.76	0.83	0.90	0.95	0.89	0.91	0.87	0.86	0.83
ResNet	0.78	0.87	0.79	0.75	**0.83**	0.71	**0.96**	0.79	0.88	0.79	0.80	0.78	0.73	**0.97**	0.87	0.94	0.87	0.82	**0.93**	0.80	0.96
DenseNet	**0.79**	**0.90**	**0.80**	0.76	**0.83**	**0.72**	0.79	**0.85**	**0.91**	**0.85**	**0.86**	**0.84**	**0.80**	**0.97**	**0.91**	**0.96**	**0.91**	**0.93**	0.89	**0.88**	0.96
SqueezeNet	0.78	0.88	0.78	0.78	0.79	0.71	0.80	0.79	0.89	0.78	0.85	0.72	0.75	0.79	0.90	**0.96**	**0.91**	**0.93**	0.88	**0.88**	**0.97**
Average	0.78	0.88	0.78	0.79	0.80	0.71	0.83	0.81	0.89	0.81	0.84	0.78	0.76	0.89	0.89	0.95	0.90	0.90	0.90	0.86	0.93



(a) AlexNet Backbone Models

(b) ResNet Backbone Models

(c) DenseNet Backbone Models

(d) SqueezeNet Backbone Models

Figure 3.5: AUC curves of different models. By using 2D digital mammogram (DM) and digital breast tomosynthesis (DBT) data in combination, the model performance can be improved significantly. DM Classifier-Use only DM data for training, DBT Classifier-Use only DBT data for training, DM-DBT Classifier-Use DM and DBT data in combination for training.

3.5 Discussion

3.5.1 Limitation

Due to the nature of medical images, most of the mammography datasets are private collections. The largest publicly available mammography dataset is the Digital

Database of Screening Mammography (DDSM), which was established in 1999, which contains 2620 cases for three classes (Negative, Benign, and Malignant). DDSM only contains digitized screen-files of mammograms and does not include any DM or DBT. Though we are using the largest known dataset in this study, the dataset size may still be considered as small compared to the nature imaging domain, such as the ImageNet dataset that has over one million images. The limited number of data may affect the model prediction power. Also, the model was trained only using the data from the same facility with machines from a single vendor. Therefore, these results must be validated across different vendors and facilities. Variability in the clinical data available in different facilities is expected. One of the future works is to collaborate with more facilities to increase the dataset size and to validate the model across different vendors, facilities, and populations.

3.5.2 Future Uses

The goal of this study is not to develop software to replace radiologists but to assist them. Radiologists are doing more than analyzing images. No computer software is expected to replace radiologists, soon. In addition, current deep learning algorithms work differently with radiologists. For example, the proposed method does not review any patient's prior images, which is another information radiologists often assess to detect breast cancer on mammograms. Other than imaging features, radiologists also consider more comprehensive factors, such as patients' risk profiles. Though deep learning models work differently with radiologists, the high performance of this model may enable it to serve as a second reader when a double reading at a mammography screening is not available. Our work emphasizes the need for using multiple imaging modalities to improve the accuracy of breast cancer classification and save experts' time on high-probability healthy patients. The proposed method is intuitive and straightforward and has a good performance. It will be easy to replicate, and we believe it will serve as a strong baseline for future research.

3.6 Conclusion

We propose a novel deep learning ensemble model for breast lesion classification, which simultaneously uses digital mammograms (DM) and digital breast tomosynthesis (DBT). We innovatively use the RankSVM algorithm on DBT to extract a fixed representation, dynamic feature image, of DBT. Dynamic feature image captures the slice-to-slice difference in DBT, which is the information often looked at by radiologists. The experiments show that when using both DM and DBT in combination, the single model performance can be improved nearly 10% on AUC and 23% on the prediction confidence. By applying an ensemble strategy on the three classifiers, the best performance can be improved to 0.97 AUC. This improvement indicates that deep learning models, like radiologists, benefit from combining both mammographic image formats. Also, the consistency of better performance across different feature extractors and classifiers suggests that our method is not limited to any specific deep learning architecture. The proposed DBT data representation method and dynamic feature image can also increase the classification performance of using DBT-only data by nearly 24%. In addition, our approach uses only the image-level labels. Due to a large number of incoming data in the daily, clinical practice, annotating images with bounding boxes is not practical. However, we believe that with more precise labels, such as bounding boxes, the performance of our model can be further improved. Our model can adapt to bounding box labeling with minor changes.

Chapter 4

Learn Fine-Grained Labels from Coarse Labeled Data

—Weakly-supervised training for abnormality detection

In the natural imaging domain, the use of deep learning methods has dramatically increased the state-of-the-art performance in object localization tasks. However, a high-performance localization network is usually trained with large training datasets with pixel-level or bounding box annotations, which are extremely costly, especially in the medical imaging domain. The prohibitively high annotation cost limited the applicability of using deep learning as the backbone for computer-aided detection (CADe) tools. We propose a novel weakly-supervised method for abnormality localization in medical images. The essential advantage of our approach is that the model only requires image-level labels and uses a self-training strategy to refine the predicted localization in a step-wise manner. We evaluated our approach on a large, clinically relevant mammogram dataset from breast cancer localization. The results show that our model significantly improves performance compared to other methods trained similarly.

4.1 Introduction

Recently, deep learning has demonstrated revolutionary potential in various medical imaging analysis tasks such as classification, localization, segmentation, image post-processing, treatment planning, etc [25, 95, 115, 166, 156, 154, 84]. In object localization tasks, fully-supervised training methods usually require a large number of training images with bounding boxes (BBs) of region-of-interests (ROIs) or pixel-level annotations [33, 107, 109, 169]. However, such fine-grained annotations are usually not available for medical images, especially for a clinically-relative breast cancer dataset [147]. Obtaining the annotations usually is expensive and time-consuming because the annotator needs months or even years of professional training. In contrast to fully supervised training, weakly-supervised training uses coarser annotations, such as image-level labels [23, 101, 12], which can significantly reduce the time and cost for annotation.

Domain adaptation is one way to train an object localization network without fine-grained labels. This approach was proposed to deal with the scenarios in which a model trained on a source distribution (dataset) is used in the context of a different (but related) target distribution (dataset) [9]. Domain adaptation for weakly-supervised localization training shows promising results in natural imaging settings [67]. To use such a method, we need firstly to train a localization network on a source dataset with fine-grained labels. After the model is trained well, we can use domain adaptation to apply the pre-trained model on a different but related target dataset without requiring fine-grained labels. However, in the medical imaging domain, source datasets with fine-grained labels are usually not available in the real world.

An attention mechanism, which usually refers to trainable attention [143, 58], can be used for weakly-supervised object localization as well. An attention map highlights the important areas of a given image. Ideally, the highlighted areas should be the ROIs of a given image. We can use the attention map for object localization. However, in practice, not all of the important areas are necessary to be ROIs. Zhang et al. [163] proposed to use self-produced guidance (SPG) masks for object localization of natural images. A SPG mask is learned from an attention mask. Each pixel in the attention mask is labeled as one out of three classes using a thresholding method. Then, the SPG masks are used as auxiliary pixel-level supervision to facilitate the training of classification networks for object localization. Their method is the-state-of-art weakly-supervised localization performance on the ILSVRC [113] dataset.

Inspired by Zhang et al. [163], we propose to use the class activation mapping (CAM) mechanism and self-training strategies to train a tumor localization network using only the image-level labels. More specifically, we use CAM heatmaps to replace the attention maps in their work.

Class Activation Mapping was originally proposed for model decision visualization [167, 119], in which the pixel values of CAM heatmaps are associated with the

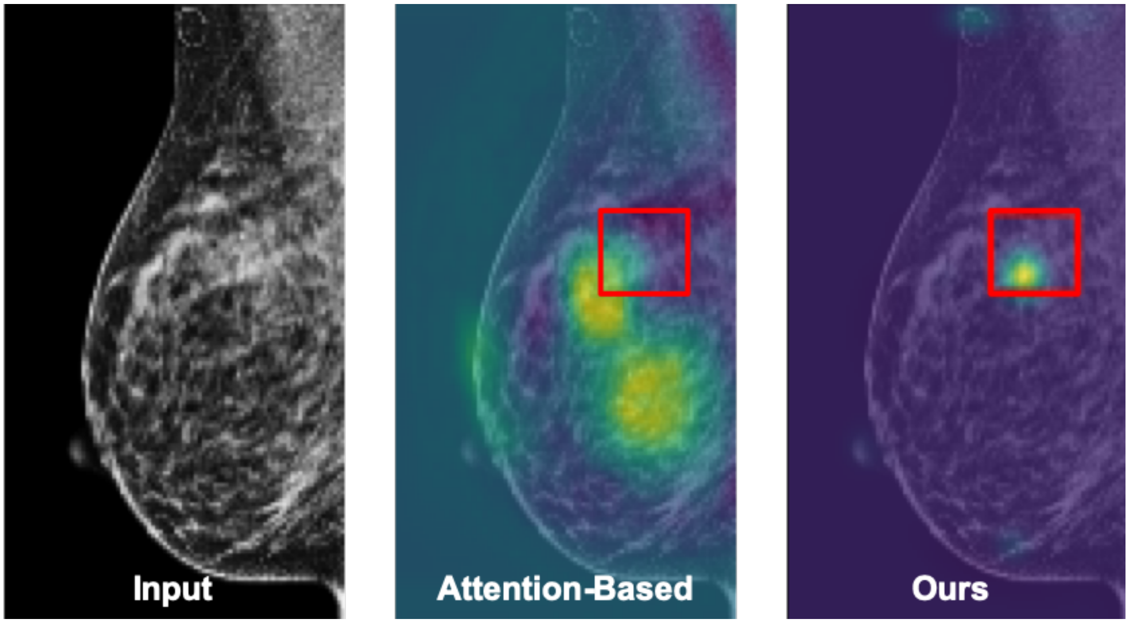


Figure 4.1: Breast tumor localization example of a given input mammogram (left), the prediction of the attention-based method (middle), and the prediction of our method (right). The red box indicates the ground truth tumor localization. The heatmap shows the predictedu location.

contribution to the classification decision. A higher value indicates a higher contribution, which implies a higher possibility of the occurrence of the object-of-interest at that location. Unlike attention maps, CAM heatmaps are only highlighting the most discriminative regions. The ROIs are usually much smaller in the medical imaging domain, and the ratio of image size to ROI size is often much higher, compared with the natural imaging domain. For instance, a typical full-field digital mammogram (FFDM) size is 3328 × 4096 pixels. However, the size of a breast tumor could be as small as 10 pixels in diameter. CAM-based methods provide a more precise localization result than the attention-based methods (Figure 4.1).

We evaluated the proposed approach on a large, clinically relevant mammogram dataset, which was recently collected from a comprehensive breast care center. Our experiment results show that the proposed method significantly improves the performance of weakly-supervised breast cancer localization tasks.

4.2 Architecture

Given an input image at the training stage, a CAM heatmap can be learned from a classification network. A trimap (a pixel-level annotation for each pixel in a CAM

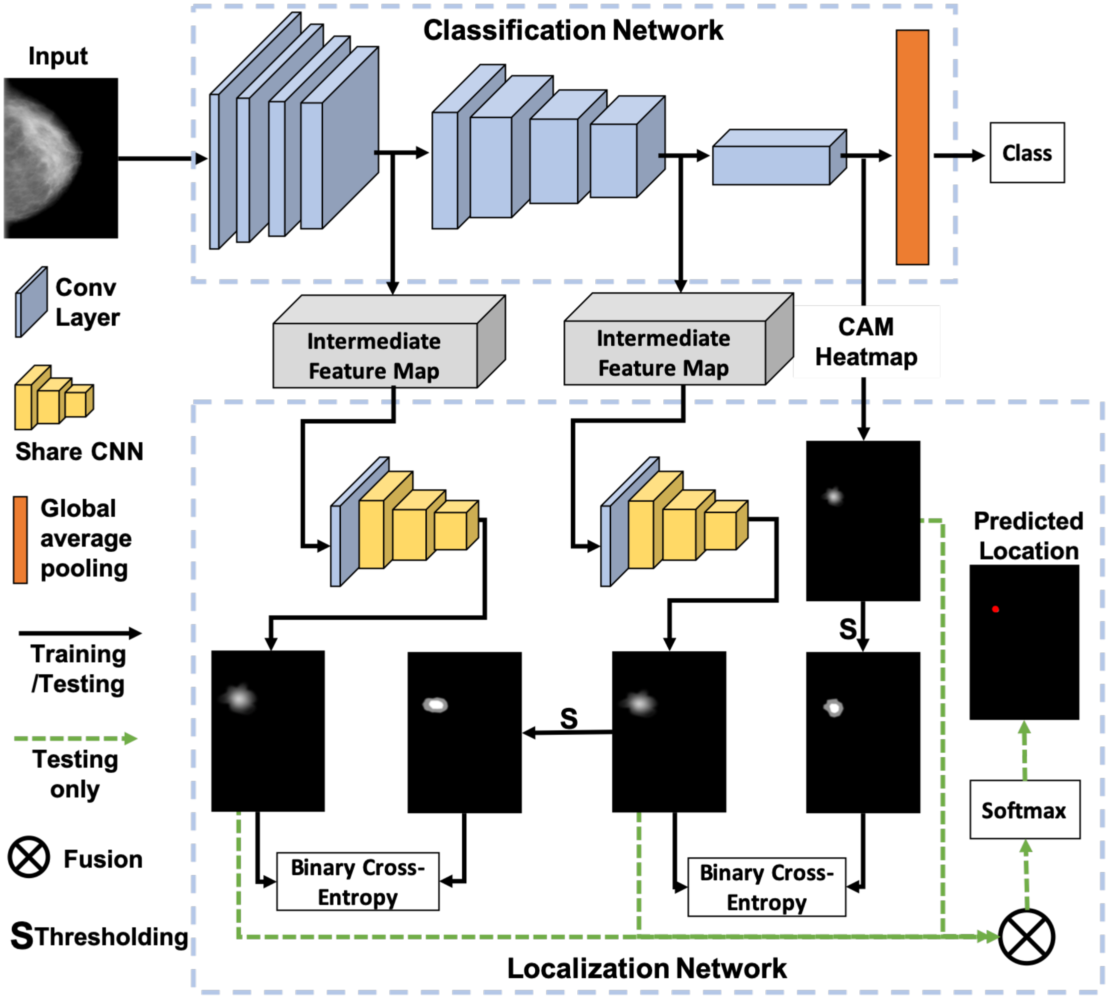


Figure 4.2: An illustration of our weakly-supervised self-training breast cancer localization model: 1) an input image passes through the classification network to extract the intermediate feature maps and CAM heatmap; 2) the localization network is trained using a self-training strategy with the intermediate feature maps and CAM heatmap; 3) at the testing stage, softmax is applied on the fused CAM and localization network outputs to find the final object location.

heatmap, in which each pixel belongs to one of three classes in the trimap) can be derived from the CAM heatmap, which highlights the high-confident foreground (ROI/tumor) pixels, the high-confident background (non-ROI/non-tumor) pixels, and the unknown pixels. Then, the trimap can be used as the pseudo-pixel-level label in a self-training convolutional neural network (CNN) localization model (Figure 4.2). More specifically, we use the foreground and background pixels in the trimap as the pseudo-pixel-level label and use the corresponding areas in an intermediate feature map (the output from a higher convolutional layer) as the input to train a CNN model for the pixel-level labeling task. The prediction of this CNN can be used to generate another pseudo-pixel-level label (trimap) to train a new CNN model that takes another intermediate feature map from an even higher convolutional layer (Conv-layer) as the input. This self-training strategy can be repeated up to K times (K equals to the number of Conv-layers in the classification model). At the testing stage, the predictions of all the self-trained CNN models were combined with the CAM heatmap. The softmax function will be applied to find the final predicted ROI localization.

4.2.1 CAM Heatmap Generalization

Class activation maps (CAM heatmaps) is generated using the global average pooling (GAP) in CNN classification networks. A CAM heatmap for a particular category indicates the discriminative image regions used by the CNN model to identify that category. More specifically, we first need to train a classification network with a GAP layer. The GAP layer follows the last Conv-layer in the network. After the GAP layer, we will have a fully-connected network follows by softmax layer, which provides the classification decision of a given image. To generate the CAM heatmap of the predicted class, we need to: 1) get all the weights connected between the fully-connected layer and the softmax class of which we want to predict. If n feature maps are presented before the GAP layer, n weights will be received. 2) We need to compute the weighted sum of the n feature maps that come from the last Conv-layer. The weighted sum generates a heatmap of a particular class. The size of the heatmap is the same as the feature map. Please see [167] for more details.

4.2.2 Self-Training

Self-training of a localization network includes two components: a pseudo-label generating strategy and a CNN model trained with the fully supervised training style. In our study, we use the self-training strategy to train multiple CNN models recursively. Each CNN model takes the output of a Conv-layer as the input and predicts a heatmap. The heatmap indicates the probability of being a tumor for each specific pixel in the input image. The pseudo-label used in the training of the base CNN model (the first model in the recursive sequence) is derived from the CAM heatmap using a thresholding method. The prediction of the base CNN model is used to generate the pseudo-label for the next CNN model, which trains in the same fashion.

More specifically, given an input image, I, we first feed it to a classification network and extract the CAM heatmap, C, and multiple intermediate feature maps, $\{F_i\}$, where $i \leq K$, K equals the number of Conv-layers in the classification model. We generate a trimap, M_C, of C using two thresholds t_f and t_b. For each pixel, p_j in C, if $p_j > t_f$, p_j is labeled as foreground; if $p_j < t_b$, p_j is labeled as background; if $t_b \leq p_j \leq t_f$, p_j is labeled as unknown. The foreground and background pixels in M_C are used to train a base CNN model (CNN_{base}).

The CNN_{base} takes F_i as the input and predicts the pixel-level label for each pixel in I. The predictions form a new heatmap, M'_C. Ideally, the high-confidence foreground and background areas in M'_C and M_C should be identical to each other. The CNN_{base} also predicts binary pixel labels of the area that was signed as unknown in M_C. A new trimap, M_{K-1}, is derived from M'_C using the same thresholding method. M_{K-1} is used to train a new CNN model, CNN_{K-1}, which takes F_{i-1} as the input and predicts M'_{K-1}. We repeat this process recursively until CNN_1 is trained, which uses F_1 as the input and M_1 as the ground truth label.

Binary cross-entropy (Equation 4.1) loss is used in all the CNN models.

$$\begin{aligned} BEC = &- \frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(\hat{y}_i)) \\ &+ (1 - y_i) \cdot \log(1 - p(\hat{y}_i)), \end{aligned} \tag{4.1}$$

where y is the label, $p(\hat{y})$ is the predicted probability of the given data point, and N is the number of data points.

4.2.3 Implementation

We used Inception-V3 as our classification network in this study. We removed all the layers after the last Inception block. Then, we added two Conv-layers of kernel size 3×3, stride 1 with 1024 kernels, a global average pooling layer, a fully-connected layer, and a softmax layer. We used the output of the last Conv-layer to compute the CAM heatmap. We extracted the outputs of the third and eighth Inception blocks as the intermediate feature maps.

The localization network contained two CNN models, one for each intermediate feature map. Each of the CNN models had three Conv-layers, followed by a sigmoid layer. The first Conv-layers of the two CNN models had 288 and 768 kernels with size of 3×3, each. The second Conv-layers of both CNNs contained 512 kernels with size of 1×1, and the third Conv-layers of both CNNs had one 1×1 kernel. The weights of the second and third layers were shared between the two CNNs.

The model was implemented in PyTorch and trained with batch size 8. The initial learning rate was 0.001. The SGD optimizer with a momentum of 0.9 was used in training. We chose the thresholds $t_f = 0.6$ and $t_b = 0.1$. We trained and tested the network on an Nvidia GTX 1080 GPU card with 8GB of memory.

4.3 Experiments

4.3.1 Dataset

We use the UKy dataset (a large, clinically related mammogram dataset) for this study. The dataset contains FFDM images for 779 positive cases and 3018 negative cases. All the mammography data were retrospectively collected from patients seen at a comprehensive breast imaging center in the United States from Jan 2014 to Dec 2017. All patients had mammograms in either craniocaudal (CC) view, mediolateral oblique (MLO) view, or both. Each image was reviewed by specialized breast radiologists. All the positive cases were proved with biopsy, and the negative cases were confirmed with more than two years of follow-up. The dataset contains cases with co-existing conditions, such as a prior benign biopsy and surgery.

The images also contain common foreign bodies, such as clips, markers, and pacemakers. The images were acquired with Hologic devices in 12-bit DICOM format at the resolution of 3328×4096 and downsampled to 832×832. Data augmentation was applied to all the positive images through a combination of reflection and rotation. Each original image was flipped horizontally and rotated by each of 90, 180, and 270 degrees. In total, 4175 positive images and 12072 negative images are used in the training stage. The dataset is randomly split into the training and validation sets on the patient-level with a $4:1$ ratio.

The training and validation sets were used for the classification model training. We manually annotated the ROIs of an additional 138 positive images with bounding boxes for testing. These images were held out during the training stage and only used in the testing stage.

4.3.2 Result

We used STL [52] and FCN_{WSL} [142] as the baseline models in this study. STL was specifically designed for breast cancer localization. FCN_{WSL} was an extension of [23] on medical related tasks. The CAM-based weakly-supervised training methods were used in both models.

We evaluated our model on localization AP, which has been widely used in weakly-supervised object localization tasks [23, 101, 142, 52]. We calculated localization AP

Table 4.1: Localization performances.

Model	STL	FCN_{WSL}	$Ours$
Loc. AP	0.43	0.26	0.52

in the following way: if the predicted location lies within the ground truth bounding box of the same class or within a tolerance distance (d), the example is considered as true positive; otherwise, it is a false positive prediction. In our experiment, only the positive class is considered for localization AP since there is no ROI on the negative class. We chose d to be equal to 12 pixels, which is the mean of [142, 52].

Table 4.1 shows localization AP for the three models. Our model achieves 0.52 localization AP, which surpasses STL by 20.93% (0.43 localization AP). FCN_{WSL} only achieved 0.26 localization AP, which is only 50% of our model.

Figure 4.3 shows the prediction results of our model. The testing images are on the left, and the predicted heatmaps are on the right. The red boxes are the ground truth bounding boxes of the malignant tumors, which indicates the ground truth localization. We used the center pixel of each heatmap as the final predicted localization. If the pixel lies in the ground truth bounding boxes or within 12 pixels, we consider the prediction as true positive.

The figure shows that our model is able to predict correct locations for both mass and calcification cases. The figure also demonstrates that our model has the ability to work in very challenging cases, such as cases with prior surgery history (the top left example in the figure).

4.4 Conclusion

We proposed a novel weakly-supervised breast cancer localization network. The proposed method only requires the image-level labels for training. No fine-grained annotation, such as bounding boxes or pixel-level labels, are needed in the training process. The model uses CAM to generate a pseudo-pixel-level label to train a localization network gradually in a self-training fashion. The evaluation result on a large clinically relevant mammogram dataset shows the proposed method has significantly improved the performance in object localization. We believe the proposed model is not only limited to breast tumor localization. It should be easily transferred to other medical imaging localization tasks with minor changes. We believe this work will serve as a strong baseline for future researchers.

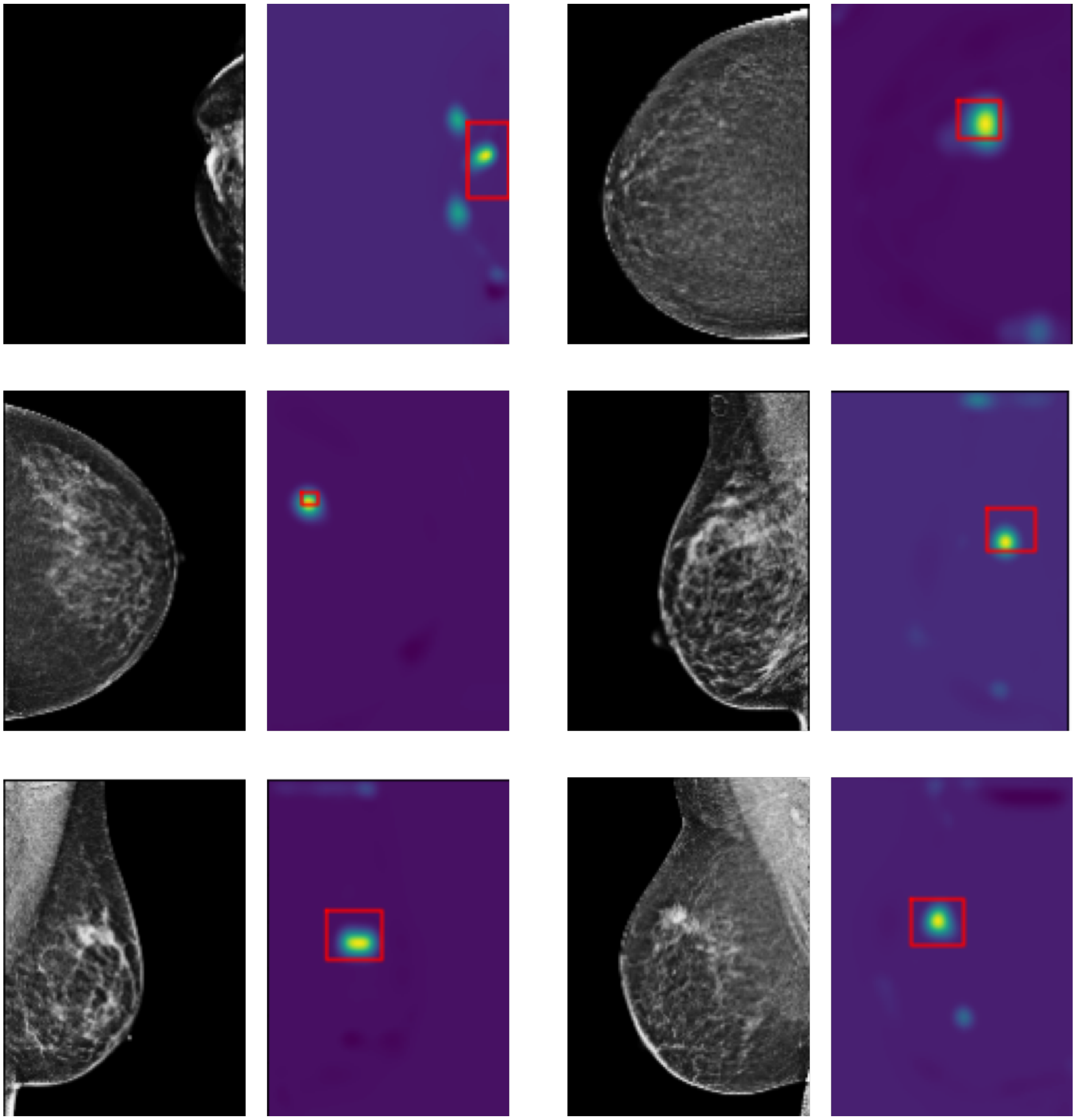


Figure 4.3: Breast tumor localization examples generated with our method. The red boxes are the ground truth bounding boxes. The heatmaps show the predicted locations.

Chapter 5

Train Deep Learning Models in Low Resource Settings

—Image feature learning under weak supervision

The number of manually annotated training instances is usually limited in the medical imaging domain due to the prohibitively high annotation cost, which posts a barrier to applying supervised deep learning techniques on medical image analysis. However, plentiful unstructured text that accompanies medical images is often readily available. The text is produced by expert physicians and includes unstructured but rich information for the corresponding images. In this work, we propose to use the unstructured text as a form of weak supervision for image feature learning through a text and image matching network. This approach is widely applicable and useful in applications where manually annotated training images are limited, but text-image pairs are readily available.

The key idea is to learn image feature representations on a large number of images that come from the same imaging domain via a network that matches imagery with text. Then various models for a downstream application (e.g., classification, detection, segmentation) can be trained using the available annotations. We evaluate the proposed method on three datasets for classification tasks and find consistent performance improvements. The biggest gains are realized when fewer manually labeled examples are available. In some cases, our method achieves the same performance as the baseline even when using 70%–98% fewer labeled examples.

5.1 Introduction

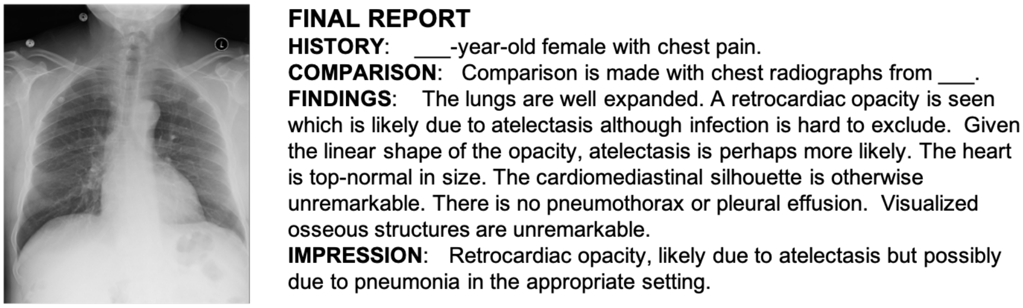
Though millions of medical images are acquired each year, few are annotated with a discrete label or a pixel-level label due to the prohibitively high cost of manual annotation [152]. Deep learning models are typically trained using large quantities of annotated data [25, 42, 26, 81, 164]. The limited number of manual annotations in the medical domain poses a barrier to applying deep learning to medical image analysis [148, 85, 147]. Much more common than manual annotation is the physicians' textual assessments and reports which typically accompany medical imagery. This unstructured text data provides rich information about the findings observed in each image, but it is difficult to directly integrate it into deep neural network training. We propose using unstructured text as a form of weak supervision for image feature learning through a text and image matching network that we call TIMNet (Text-Image Matching Network). The network learns image feature representations from a large number of text and image pairs in a weakly-supervised fashion. Then, various models for a downstream application can be built based on the learned feature representations with only a small labeled dataset.

The proposed method is widely applicable, but our focus is on the medical imaging domain, in which textual findings are often readily available but obtaining manual image-level labels is expensive. Our key contribution is in leveraging these matched text-image pairs to train a two-branch neural network that can help build effective models for downstream predictions relying only on image input. We demonstrate the proposed method on the MIMIC-CXR [62], MendeleyV2 [65], and Kather5000 [64] datasets for binary, multi-class, and multi-label classifications. In addition, we also investigate the transferability of the learned feature representations between datasets and imaging modalities. Our experiments show that the proposed method significantly reduces the need for manually annotated data by up to 98%.

5.2 Background

5.2.1 Medical Imaging Annotation Availability

Over the years, the number of acquired medical image datasets has increased dramatically. However, the number of images that can be used effectively for machine learning (incl. deep learning) remains quite small due to the limited availability of manual labels [152]. For instance, the cancer imaging archive (TCIA) [7] hosts one of



FINAL REPORT
HISTORY: ___-year-old female with chest pain.
COMPARISON: Comparison is made with chest radiographs from ___.
FINDINGS: The lungs are well expanded. A retrocardiac opacity is seen which is likely due to atelectasis although infection is hard to exclude. Given the linear shape of the opacity, atelectasis is perhaps more likely. The heart is top-normal in size. The cardiomediastinal silhouette is otherwise unremarkable. There is no pneumothorax or pleural effusion. Visualized osseous structures are unremarkable.
IMPRESSION: Retrocardiac opacity, likely due to atelectasis but possibly due to pneumonia in the appropriate setting.

Figure 5.1: An example of a chest x-ray (left) with the radiology report (right) from the MIMIC-CXR dataset.

the largest collections of cancer images that are accessible for public download; while the archive includes 123 datasets as of August 1, 2020, only four of them contain more than 1000 cases, and a majority have fewer than a hundred cases.

On the other hand, almost every clinical imaging procedure/event is accompanied by at least one clinician authored textual report summarizing the corresponding findings from the medical images generated. These reports are a required artifact of the typical care delivery paradigm. They inform further patient care and are part of the archived record of the patient's interactions with the healthcare system. For instance the MIMIC-CXR dataset contains 227,835 radiographic studies of 64,588 patients with 368,948 chest X-rays and associated radiology reports (Figure 5.1). These imaging reports are rich with information about the corresponding images, but their use in training deep models for image classification is non-trivial. Although image-level labels may be derived from the reports using machine learning techniques, such as in NegBio [104] or CheXpert [55], the accuracy of the derived labels could still be problematic. In addition, labeled data may still be needed to train a machine learning model that derives labels from unstructured text.

In this study, instead of trying to infer labels from the reports, we use the reports directly by letting the model directly glean the associations between text and images. Intuitively speaking, our setup transfers the strong expert generated language signal to the image feature generation part of the architecture, priming it more powerfully for downstream training on supervised tasks.

5.2.2 Weakly-Supervised Learning

Neural architectures are usually data-hungry, often relying on large labeled datasets, posing a frequent challenge for the medical imaging domain [148, 85]. Weak supervision methods have been proposed to improve the performance of deep net models when dealing with smaller training datasets [168, 139]. In general, weak supervision may be categorized into two groups: 1) incomplete supervision and 2) inexact supervision.

Incomplete supervision arises when only a small portion of a given dataset is labeled. Active learning methods [120, 71, 157] may be useful when the labeling is incomplete. These methods try to select the most valuable unlabeled instance to hand-annotate based on specific criteria of informativeness. However, when only a few labeled examples are available, the performance of an active learning model may be unstable. Semi-supervised learning [170, 10, 60, 16] is another alternative to incorporate incomplete supervision. These methods are based on assumptions that link the input distribution to the decision properties. Some techniques [38, 8, 66] propagate labels from the labeled to unlabeled data, and use the larger, newly labeled data for training. They assume that the high confidence predictions are correct, and hence proceed to use them in the training process. The approach uses such assumption is also called self-supervised learning approach.

Inexact supervision applies to scenarios where the given supervision is not as exact as desired. A specific example is when we only have image-level labels to train a segmentation model. Class activation mapping (CAM) [167, 119] is widely used when training a segmentation network with only coarse-grained labels. CAM was initially proposed for model decision visualization, in which the pixel values are associated with the contribution to the classification decision, and can be used to generate saliency maps [23, 56]. Multi-instance learning (MIL) is a more general approach when dealing with inexact supervision. MIL has been applied to different tasks, such as classification [17] and segmentation [59]. This approach takes a set of labeled bags containing many instances as training data. An instance in the bag is a small, unlabeled image patch that can be extracted from an image. At test time, one needs to convert the test image into patches and predict each patch [160, 144].

In this study, given an image-text pair, we exploit latent connections between linguistic artifacts in textual findings and image features as a form of inexact supervision. These connections indirectly guide and pre-train an effective image feature extractor, which can be fine-tuned for various downstream tasks. The model thus learns more discriminative feature representations without acquiring additional labels.

5.2.3 Text and Image Matching

Text-image matching has become a popular research topic in recent years. Broadly, the existing matching methods can be divided into two categories: 1) local representation matching and 2) global representation matching.

Local representation matching focuses on local feature matching between text and images. For instance, we can extract objects from images and then match the objects to words in a given text [63, 79, 145, 86, 146]. One drawback of the local representation matching method is that it usually relies on pre-trained object detectors, such as R-CNNs [34, 33, 108]. However, a pre-trained, high-performance object detector is usually not available in the medical imaging domain, given the complexities of the tasks.

The global representation matching method usually contains three procedures: 1) image feature embedding, 2) text feature embedding, and 3) measures of distance between the two embeddings. For instance, Kiros et al. [69] use a convolutional neural network (CNN) to encode images and a long short-term memory network (LSTM) [50] to encode the full text. The triplet loss is used to pull the embeddings of the matched text and image closer to each other and push the unmatched ones further. Wehrmann et al. [149] encode the text data using an efficient character-level inception module that convolves over characters in the text. Sarafianos et al. [116] train a text-image matching network by using a ResNet101 [44] network for imaging processing and a transformer-based model [21] with an LSTM based encoder for text processing. Prior efforts seem to focus on the matching problem from the perspective of text-based querying systems for image retrieval. In contrast, our goal is to imbue an image classifier with the knowledge obtained from learning to match a medical image with the associated clinician generated textual report. We use the global matching approach with a two-branch setup one each for image processing and text processing. The absolute difference between the two output feature vectors is fed to a classification network, which predicts if the input image and text are a valid pair — the text snippet is in fact the correct findings report for the image.

5.2.4 Natural Language Encoding

The deep learning revolution in natural language processing (NLP) started with the idea of representing words as dense embeddings in a real vector space as opposed to the typical multi-hot encoding. Self-supervised pre-training of neural word embeddings picked up with the now-famous Word2Vec [97, 96] approach. The main idea is based on the distributional hypothesis that similar words tend to share distributional similarities appearing in similar or at times in the same contexts. However, the drawback of this method is the embedding is independent from the context in which the word appears. Thus, this method cannot address polysemy or homonymy. Nevertheless, word2vec based embeddings pre-trained on large corpora (e.g., Wikipedia) have been used to initialize deep neural networks and are subsequently fine-tuned in a supervised task.

Modern NLP has moved toward more contextualized embeddings by using a language modeling based objective for self-supervision instead of the simpler distributional one used by word2vec. The main idea is to predict words based on prefixes in longer snippets of text and recently predict even masked words in the middle of a text snippet. The transformer [143] architecture uses an encoder-decoder structure and attention mechanism to translate one sequence to another sequence. It can encode the contextual information of a word from distant parts of a sentence. Unlike a conventional recurrent neural network (RNN), a transformer does not require the sequential data to be processed in order leading to new practical gains in efficiency.

The *bidirectional encoder representations from transformers* (BERT) [21] architecture trains a transformer by jointly conditioning on both left and right contexts in all layers. A pre-trained BERT model can be fine-tuned to create state-of-the-art models for various tasks without substantial task-specific architectural modifications [134, 78]. In this study, we use a pre-trained BERT model as the text processing branch that encodes a given text into a feature vector.

5.3 Method

Our proposed model TIMNet consists of two modules: 1) a weakly-supervised image feature learning module via a text-image matching network and 2) a downstream application module that is designed for a specific task. A CNN image feature extractor is shared between the modules (Figure 5.2). The two modules can be optimized in an interleaved manner or trained in separate phases. For simplicity, we present the two modules in a two-phase training setup.

5.3.1 Weakly-Supervised Image Feature Learning

Textual annotation based weak-supervision is carried out in TIMNet through a two-branch network, one for text processing and the other for imaging processing. The network takes a text and image pair as input and predicts whether they are naturally

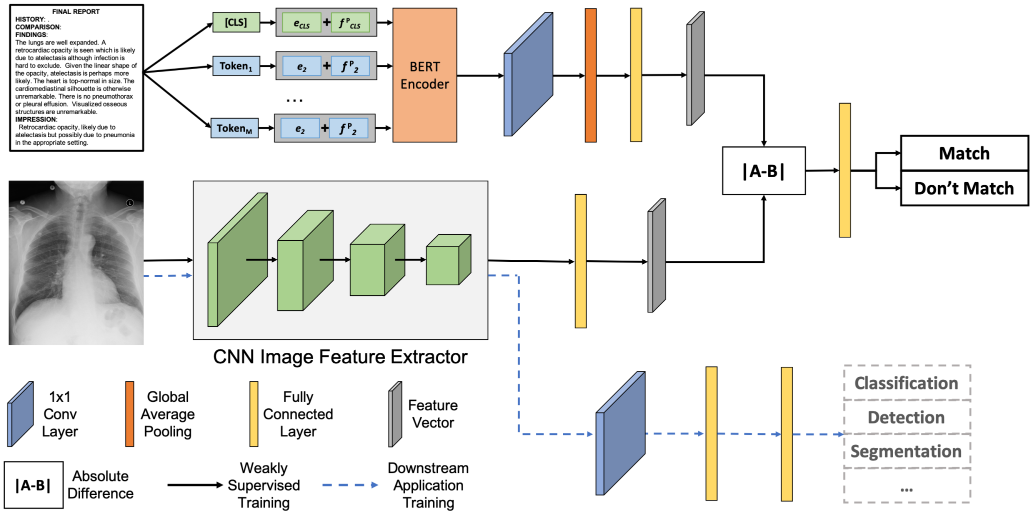

FINAL REPORT
HISTORY: .
COMPARISON:
FINDINGS:
The lungs are well expanded. A retrocardiac opacity is seen which is likely due to atelectasis although infection is hard to exclude. Given the linear shape of the opacity, atelectasis is perhaps more likely. The heart is top-normal in size. The cardiomediastinal silhouette is otherwise unremarkable. There is no pneumothorax or pleural effusion. Visualized osseous structures are unremarkable.
IMPRESSION:
Retrocardiac opacity, likely due to atelectasis but possibly due to pneumonia in the appropriate setting.
[CLS]
Token1
TokenM
eCLS + fPCLS
e2 + fP2
e2 + fP2
...
BERT Encoder
|A-B|
Match
Don't Match
CNN Image Feature Extractor
Classification
Detection
Segmentation
...
1x1 Conv Layer
Global Average Pooling
Fully Connected Layer
Feature Vector
|A-B|
Absolute Difference
Weakly Supervised Training
Downstream Application Training

Figure 5.2: The TIMNet matching architecture. 1) Weakly-supervised image feature learning through a text-image matching network (solid black line). 2) Downstream application training using a small dataset (dashed blue line).

related. For instance, whether the text is the radiology report that corresponds to the radiographic image. We use a pre-trained BERT model as the backbone of the text processing branch and a CNN model for the image processing branch.

The pre-trained BERT model is used as a text feature extractor that encodes input text as a vector. The output of the BERT model is then passed through a 1×1 convolutional (Conv) layer, which translates it to the target domain. After that, global average pooling (GAP) is applied to all the hidden states. A fully-connected (FC) layer is added to transfer the pooled hidden states to a feature vector (denoted as V_t). The CNN model of the image processing branch contains a feature extractor (i.e., multiple Conv layers) and an FC layer. The feature extractor predicts a block of features from input images. The FC layer converts the image features to a feature vector (denoted as V_i). Finally, the absolute difference between the outputs of the two branches ($|V_t - V_i|$) is fed into a shallow classification network that predicts whether the pair of feature vectors belong to the same example or not.

The input of the two-branch network is a text and image pair with a label indicating whether the text-image pair corresponds to the same imaging event or not. A true pair means the text is the correct report for the corresponding image; otherwise, it is a false or fake pair, where the text is randomly selected from other reports in the dataset. We ensure that the numbers of true pairs and false pairs are balanced in training. The length of each piece of text is pre-processed to 256 words at the word embedding stage. We add 0s for the texts shorter than 256 words, and we snip much longer texts to 256 words. The output of this first weak-supervision phase is a probability estimate of the input text-image pair being a true match.

5.3.2 Fine-Tuning for Downstream Tasks

At the end of the weak-supervision phase described earlier, the hypothesis is that the matching process has pre-trained the image feature learning component parameters to an extent that it needs fewer supervised instances for a downstream image-related task. This in turn relies on our high-level intuition that there is nontrivial transferable signal available in the textual annotations to improve imaging tasks down the line.

The fine-tuning of the image-processing branch in TIMNet for downstream tasks is fairly straightforward. Although it can be done for a variety of applications, in this study, we demonstrate the proposed method for classification tasks. To build a downstream application model, we can either add additional Conv layers and FC layers to the image processing branch, or use it as-is by retraining the FC layers. Since we only need to optimize the few additional layers from scratch, while the rest of the network is already pre-trained on a larger set of images from the same domain, the total number of required training instances for fine-tuning could be much smaller than training the entire network from scratch.

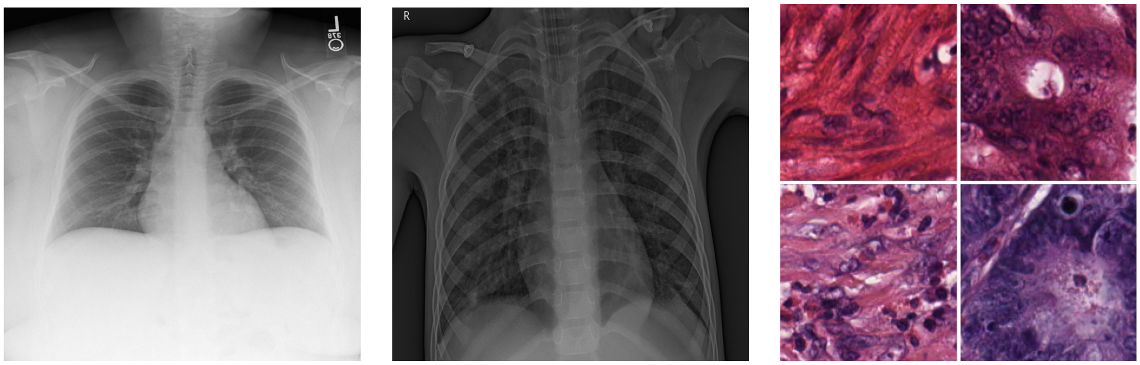
5.4 Evaluations and Discussion

5.4.1 Experiment Setup

Implementation

We implement the experiments in Pytorch [103]. We use a pre-trained BERT model (specifically, `bert-base-uncased` from HuggingFace [153]) as the backbone of the text processing branch in TIMNet. A 1×1 Conv layer, a GAP layer, and an FC layer with 512 neurons are added to the BERT model to process its hidden outputs for each word. We use the ResNet-18 [44] model as the backbone of the imaging processing branch in TIMNet. A 1×1 Conv layer and an FC layer with 512 neurons are added before and after the GAP layer, respectively.

All the Conv layers of the image processing branch in TIMNet are used as a feature extractor in the downstream networks. A shallow CNN classification network is added on top of the feature extractor. The CNN classification network contains a 1×1 Conv layer, an FC layer with 512 neurons, and an output layer with various



(a) A chest X-ray from MIMIC-CXR

(b) A pediatric chest X-ray from Mendeley-V2

(c) Four histological patches from Kather5000

Figure 5.3: Examples from different datasets used in evaluation.

numbers of neurons for different tasks. Cross-entropy loss and Adam optimizer [68] with a learning rate of 0.0001 are used for both weak-supervision (phase one) and downstream application training (phase two).

Training

TIMNet is pre-trained for 50 epochs using the text and image pairs from the training set of MIMIC-CXR. The "findings" portion of the radiology reports is used as the text input. The downstream networks are trained using varying amounts of the training data and image-level labels from the MIMIC-CXR, Mendeley V2, and Kather500 datasets—ranging from 0.5% to 100%. Each downstream model is trained for 100 epochs with batch size 16 for five trials. The results reported in this paper are the averaged performances over all five trials.

5.4.2 Datasets

We use the MIMIC-CXR, MendeleyV2, and Kather5000 datasets in this study. The MIMIC-CXR dataset is used in training of both TIMNet pre-training and downstream applications training. MendeleyV2, and Kather5000 are only used for downstream application training.

MIMIC-CXR

The dataset contains 227,835 radiographic studies of 64,588 patients with 368,948 chest X-rays and the associated radiology reports (Figure 5.1 and Figure 5.3a). In the official train/validation/test split, the validation set and test set are small with

2,991 and 5,159 images, respectively. We combine the official validation and test sets together to form our own validation set. The chest X-ray images are resized to 500 × 500. We use the official train set and our validation set in both TIMNet pre-training and downstream application training.

Mendeley-V2

The dataset (Figure 5.3b) is a pediatric chest X-ray dataset that includes 4,273 pneumonia images and 1,583 normal images. Though the imaging modality is the same with MIMIC-CXR, the patient demographics are different. In addition, the images in Mendeley-V2 also have different appearances. We used the original train/validation split of this dataset to train a downstream application model that evaluates the transferability of pre-trained weights between different datasets while retaining the same imaging modality.

Kather5000

This dataset (Figure 5.3c) contains 5,000 histological images of 150 × 150 pixels. Each image belongs to exactly one of eight balanced categories: tumor epithelium, simple stroma, complex stroma, immune cells, debris, normal mucosal glands, adipose tissue, and background (no tissue). All images are RGB, $0.495\mu m$ per pixel, digitized with an Aperio ScanScope (Aperio/Leica Biosystems), magnification 20×. We randomly partition the dataset into training and validation sets with a 4:1 ratio. The dataset is used to train a downstream application model that aims to evaluate the transferability of pre-trained weights between different imaging modalities.

5.4.3 Evaluation Method

We evaluate the proposed method through downstream application performance and the degree of need for labeled instances for supervision. We denote the downstream models with TIMNet pre-trained weights as *Ours* and models without the pre-trained weights as *Base*. The weights of *Base* models are randomly initialized, and the weights of *Ours* variants are pre-trained using TIMNet. All compared models have the same architecture.

We use (a). accuracy (ACC), the area under the receiver operating characteristics curve (auROC), F1 score (F1), and average precision (AP) as the evaluation metrics for binary classification tasks, (b). auROC and AP for multi-label classification tasks, and (c). ACC, F1, Prec, and recall for multi-class tasks.

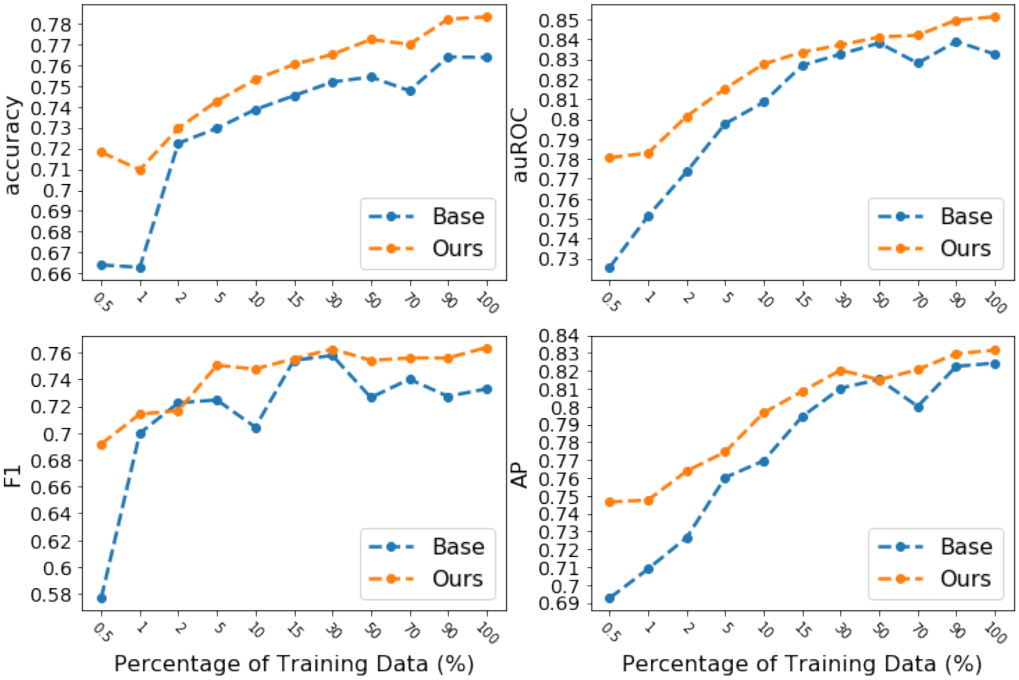


Figure 5.4: Binary classification results on MIMIC-CXR dataset.

5.4.4 Classification on MIMIC-CXR

We first present the evaluation results of downstream applications that are trained using the MIMIC-CXR dataset. The text-image pairs in this dataset are also used in TIMNet pre-training. Two downstream application models are tested, a binary classification model and a multi-label classification model. The MIMIC-CXR dataset contains 14 labels (with 13 labels for different abnormalities and one label for the normal case), which are derived from the radiology reports using NLP tools. The binary classification model predicts whether an abnormality exists in an image, and the multi-label classification model predicts what kind of abnormality exists in an image. The output of the binary classification model is Boolean, and the output for the multi-label task is a multi-hot vector, with 1 indicating the presence of a particular class. Multiple classes can present within the same image at once.

Figure 5.4 shows the result of the binary classification task on the MIMIC-CXR dataset. The results reveal that the proposed model has superior performance compared with the *Base* model in all settings, with better gains when only few labeled images are available. For instance, when using 0.5% of the labeled data (≈ 1,850 instances), the *Base* model has an accuracy of 66.41%, while the *Ours* has a 71.81% accuracy. The highest accuracy of *Base* is 76.38%, when it is trained with 90% of the labeled data (≈ 339,340 instances). *Ours* surpasses the best performance (across all

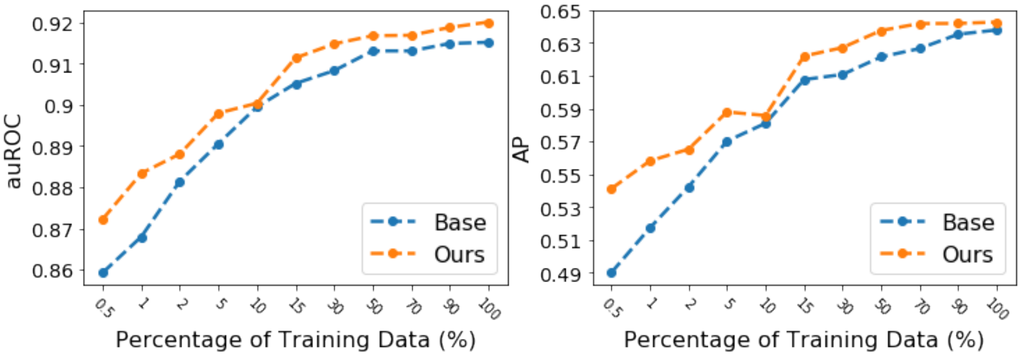


Figure 5.5: Multi-label leanring results on MIMIC-CXR dataset.

metrics) of *Base* with only 30% of training data. Thus the need for manual labels is reduced by 70% when using the proposed method.

Figure 5.5 for multi-label classification results also show the superior performance of *Ours* compared to the *Base* models. As in the binary case, TIMNet is able to significantly reduce the need for manual labels to achieve comparable performance. The *Base* model reaches its best performance (0.9152 auROC) with 100% of training data, while *Ours* can achieve a similar performance (0.9148 auROC) with only 30% of training data.

5.4.5 Transferability of Pre-Trained Weights

The feature extractor used in *Ours* is pre-trained via TIMNet on MIMIC-CXR, a chest X-ray dataset. In this section, we evaluate the transferability of these pre-trained weights between different datasets and imaging modalities. More specifically, we first evaluate the model performance on Mendeley-V2, which is also a chest X-ray dataset, but with different patient demographics compared with MIMIC-CXR. Then, we evaluate the proposed method on Kather5000, a dataset of a different imaging modality.

Different dataset, same modality

Figure 5.6 shows the model performance for pneumonia classification on the Mendeley-V2 dataset of pediatric chest X-ray images. From the results, we can see that TIMNet's pre-trained features work surprisingly well as it is able to reduce the need for labeled data by 98.33% as elaborated next. The *Base* model achieves its highest accuracy of 87.52% using 30% of the training data, and the highest auROC of 0.9333 also using 30% of the training data, while *Ours* outperforms *Base* with using only 0.5%

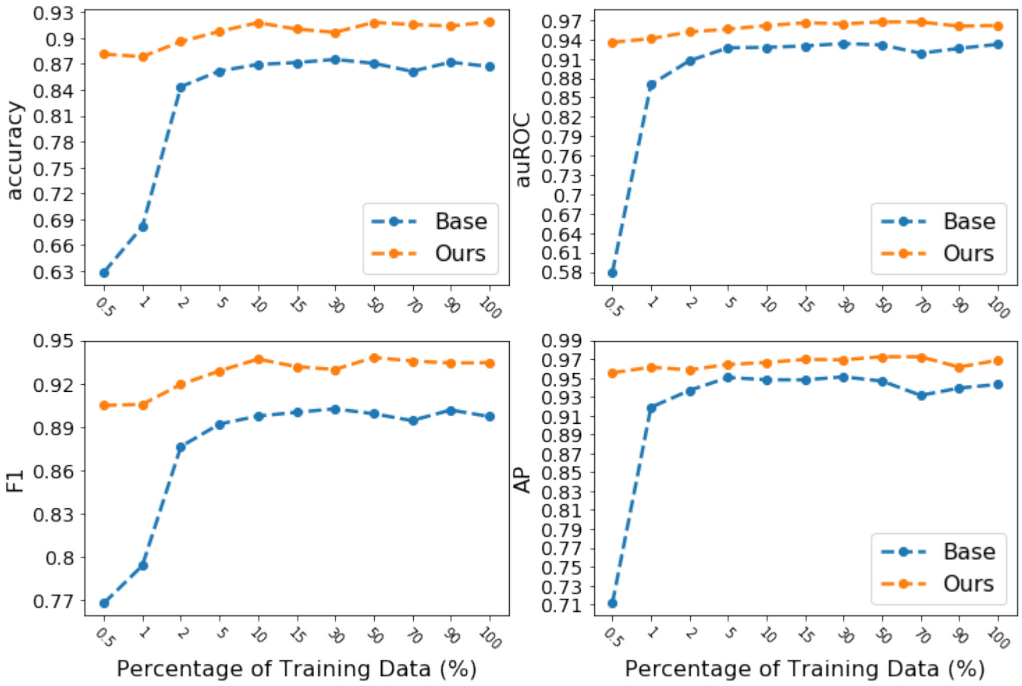


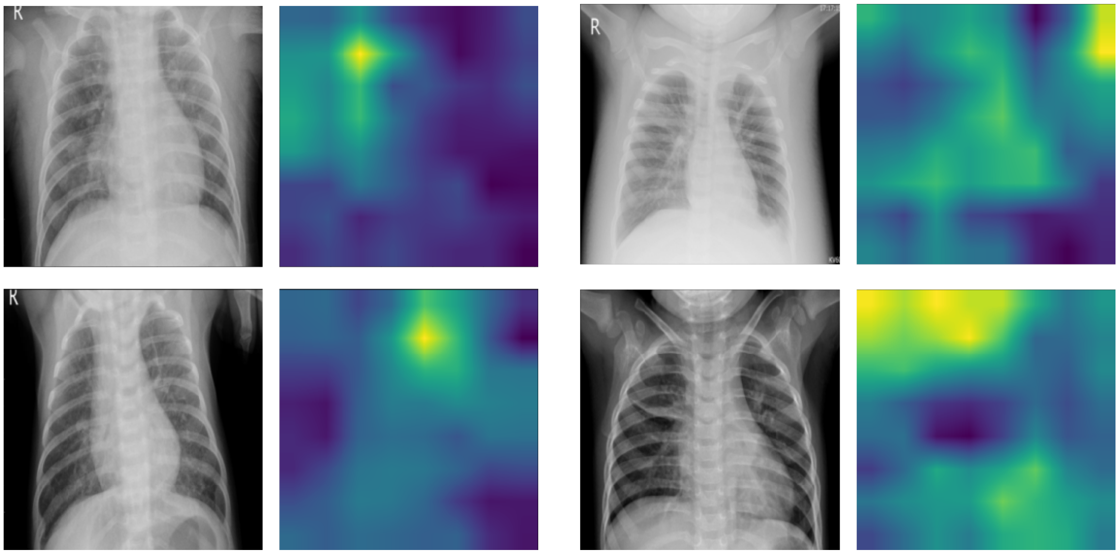
Figure 5.6: Binary classification results on Mendeley-V2 dataset.

of training data with an 88.14% accuracy and a 0.9352 auROC. Thus the reduction in training instances is $(30 - 0.5)/30 = 98.33\%$. The highest performances of *Ours* are 91.87% accuracy and 0.9613 auROC. These are also nearly 5% (ACC) and 3% (auROC) higher than the *Base* model.

Figure 5.7 shows four class activation mapping (CAM) [167] visualizations of the proposed method. The pixel values in CAMs are associated with the contribution to the classification decision. A higher value (brighter color) indicates a higher contribution to the class decision. All four cases in the figure are ground truth positive cases. The CAMs reveal that the model focuses more on the upper chest areas for the correct cases (Figure 5.7a). In the X-rays, we can see that the corresponding areas show some concerns about pneumonia. However, for the incorrect cases (Figure 5.7b), the model appears to focus on the edges of the images, which are not meaningful areas to look at because the areas are either outside of a human body or outside of the chest area.

Different dataset, different modality

Figure 5.8 shows the multi-class classification performance on Kather5000, a histological imaging dataset. The imaging modality in this dataset is very different from chest X-rays. While chest X-rays are grayscale images with only one channel, histological



(a) Two true positive predictions

(b) Two false negative predictions

Figure 5.7: Four CAM visualizations for pediatric chest X-ray pneumonia classification on the Mendeley-V2 dataset with chest X-ray on the left and CAM on the right. Top: Two true positive predictions. Bottom: Two false negative predictions.

Table 5.1: Text and image matching results

Dataset	Accuracy	auROC	F1 Score	Prec	Recall	AP
MIMIC-CXR	0.74	0.83	0.74	0.67	0.82	0.77

images are in color with three channels. The results reveal that the proposed method can reduce the need for labeled training data by 70% on this dataset. The *Base* model achieves the best performance with 95.85% accuracy using 100% of the training data, while *Ours* achieves a similar performance using only 30% of the training data. With 100% of the training data, we can push the model performance to 97.08% accuracy.

5.4.6 Text and Image Matching Result

At the text-image matching stage, TIMNet takes a text-image pair as input and predicts whether the text and the image constitute a true pairing (corresponding to the same imaging event). When testing TIMNet, we randomly feed a true or a false pair to TIMNet from a balanced set of such pairs. The pre-trained TIMNet used in this study has a text-image matching performance of 74% accuracy and 0.83 auROC. Figure 5.9 shows the auROC curve and the area under the precision-recall curve (auPRC) of this model. Table 5.1 shows the all pertinent results of this evaluation.

Figure 5.10 shows four CAM visualizations of TIMNet on the text-image matching

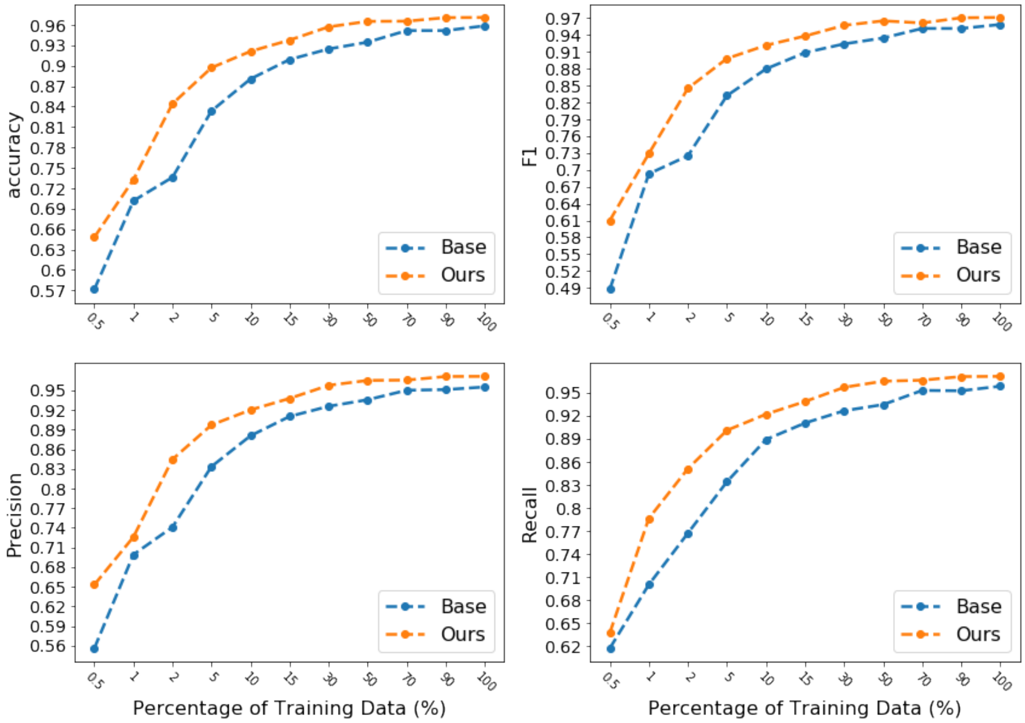


Figure 5.8: Multi-class classification results on Kather5000 dataset.

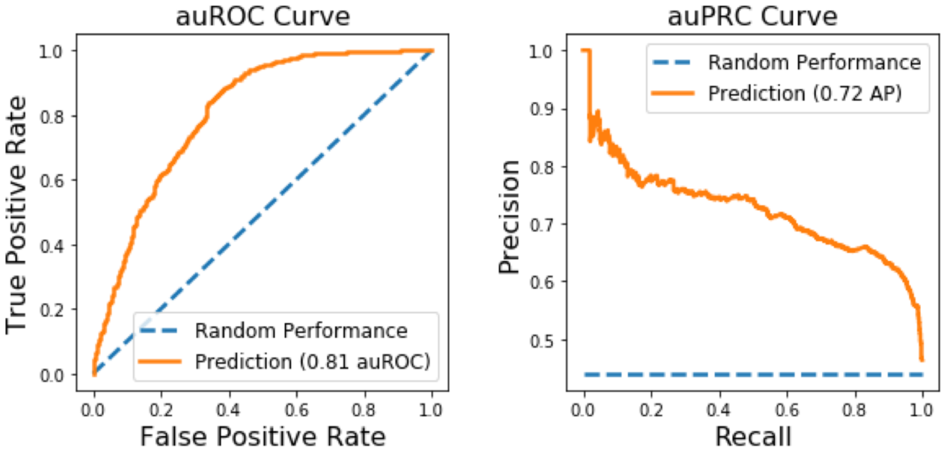
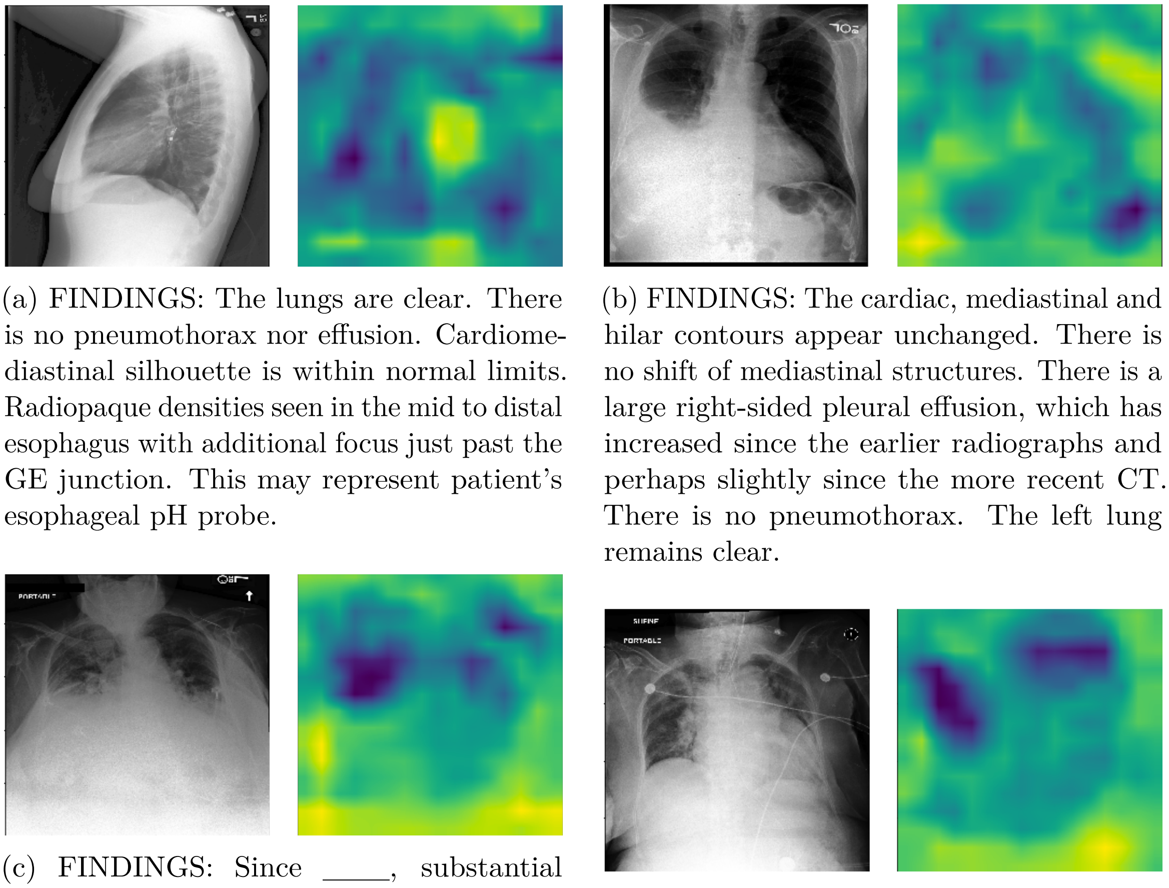


Figure 5.9: Text and image matching results on the MIMIC-CXR.

task. The findings portions of the radiology reports are displayed below the images. The CAMs suggest that the decisions made by TIMNet are reasonable. For instance, in Figure 5.10a, the radiology report mentions radiopaque densities in the mid to distal esophagus, and the CAM appears to show that in the middle part of the image. For Figure 5.10b, the radiology report indicates increased right-sided pleural effusion, and CAM shows more significant contributions near the effusion areas on the right-hand



(a) FINDINGS: The lungs are clear. There is no pneumothorax nor effusion. Cardiomediastinal silhouette is within normal limits. Radiopaque densities seen in the mid to distal esophagus with additional focus just past the GE junction. This may represent patient's esophageal pH probe.

(b) FINDINGS: The cardiac, mediastinal and hilar contours appear unchanged. There is no shift of mediastinal structures. There is a large right-sided pleural effusion, which has increased since the earlier radiographs and perhaps slightly since the more recent CT. There is no pneumothorax. The left lung remains clear.

(c) FINDINGS: Since _____, substantial pulmonary edema is increased, bilateral layering pleural effusions, right greater than left, are increased with persistent bibasilar and retrocardiac atelectasis. Lung volumes remain low. Cardiomegaly is difficult to evaluate but also appears worse. No pneumothorax.

(d) FINDINGS: ET tube is seen with tip approximately 1.8 cm from the carina. Enteric tube seen passing below the inferior field of view. Lower lung volumes are noted on the current exam with bilateral parenchymal opacities which could be due to edema or infection. Prominence of the right hilum is again noted. Moderate cardiomegaly and appears to have progressed since prior could potentially bein part due to changes in positioning. No acute osseous abnormalities. Surgical clips project over the left chest wall/axilla.

Figure 5.10: Examples of CAM visualizations of text and image matching on MIMIC-CXR.

side of the figure. We can also see a similar correspondence between CAM based image segment contributions to the model decision and the textual report in Figure 5.10c and Figure 5.10d.

5.5 Conclusion

The main objective of our work is to demonstrate the potential of clinician authored textual findings that accompany most medical images in improving image-related supervised ML applications. Such textual narratives are readily available and routinely curated as part of healthcare operations and hence are a natural resource to leverage. The central premise on which our effort stands is the insight that in the latent neural dense vector representation space, it may be viable to transfer linguistic signals that characterize expert summaries of images to downstream image-based tasks through weak-supervision. Based on the experiments and evaluations in this paper, we believe we have successfully verified this insight for classification tasks.

At the core of our methodology is a two-branch architecture, TIMNet, that identifies if a textual finding corresponds to the supplied image. Subsequently, the image branch is further fine-tuned for downstream supervised tasks. The improvements are substantial in that small fractions (2%–30%) of the available full training data are needed to achieve the same performance as baseline models that do not exploit textual findings. Additionally, the benefits also persist across datasets and modalities, which is an excellent affordance when transferring signals from models learned on deidentified textual findings (e.g., MIMIC-CXR) to other classification settings that either have fewer or no textual annotations (due to HIPAA and other privacy restrictions).

During our experiments, we discovered that better text-image matching performances usually lead to improved downstream application performances in terms of higher accuracy and need for fewer labeled images. Thus, one of our future directions will be to further innovate on the matching framework to improve the associated performance, to more tightly couple the textual features and image representations. Another important future direction is to see if our text-image matching setup can actually transfer the image signal to downstream tasks in the NLP domain for clinical text. We believe this bidirectional feedback may help in extracting named entities (e.g., drugs, comorbidities, anatomical sites) and relations connecting such entities (e.g., adverse drug reactions) from a variety of notes. These information extraction tasks are also usually plagued by a lack of large training datasets (esp. public ones) due to stricter regulatory constraints governing textual data in medicine. Overall, we hope our work spurs further intere

Chapter 6

A Necessary Step to Safe AI

—Deep learning model calibration

“Who will take responsibility if the result of the AI system was wrong in clinical practice?" is a question that has been asked repeatedly by medical experts. As the question pointed out that liability is one of the potential issues preventing adopting state-of-the-art deep learning techniques in the medical field. In the medical imaging analysis domain, many works have shown that modern neural networks can achieve super-human performance in a wide range of tasks. However, these works have primarily focused on accuracy, ignoring the important role of uncertainty quantification in medical decision-making. Empirically, neural networks are often miscalibrated and dramatically overconfident in their predictions. This miscalibration could be problematic in any automatic decision-making system, but we focus on the medical field in which neural network miscalibration has the potential to lead to significant treatment errors.

In this chapter, we propose a novel approach to neural network calibration that maintains the overall classification accuracy while significantly improving model calibration. The proposed approach is based on ECE, which is a standard metric for quantifying model calibration error. As such, it is a natural and empirical way of assessing model calibration. Our approach can be easily integrated into any classification task as an auxiliary loss term, thus not requiring an explicit training round for calibration. We show that our approach reduces calibration error significantly across various architectures and datasets and that it performs better than temperature scaling, the current state-of-the-art approach.

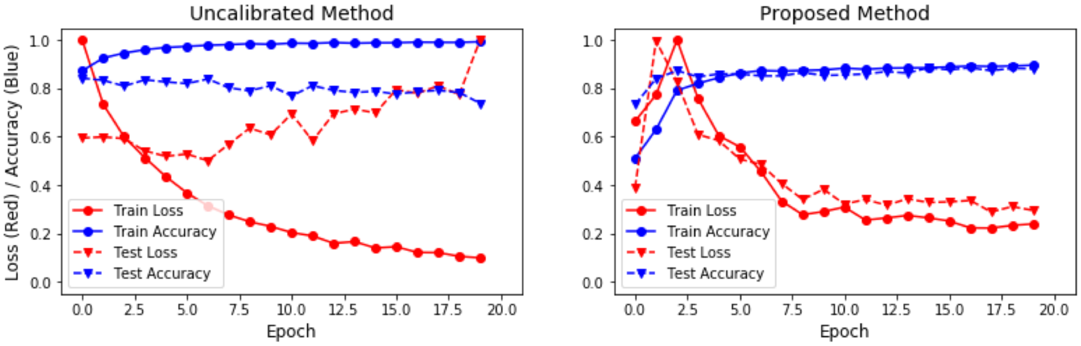


Figure 6.1: Left: The train loss/accuracy and test loss/accuracy of the uncalibrated model. The model is overfitted after the 7^{th} epoch, where the train loss keeps decreasing but the test loss keeps increasing. Right: The train loss/accuracy and test loss/accuracy of our method. The DCA term penalizes the model when the loss reduces but the accuracy is plateaued. Both the train and test losses maintain at the same level after the 7^{th} epoch.

6.1 Introduction

Recent advances in deep learning research have dramatically impacted the research field of medical imaging analysis [115, 135, 81]. Many high-performance deep learning models have been developed in the field [25, 156, 95, 166]. Researchers are actively pushing convolutional neural networks (CNNs) to have higher and higher accuracy, while uncertainty quantification is often ignored when evaluating these models [110, 109, 83, 154]. However, uncertainty quantification of neural networks is important, especially in automatic decision-making settings in the medical field. An automated method that achieves high accuracy, but captures uncertainty inaccurately, such as providing inaccurate confidence or probability of a specific prediction, could lead to significant treatment errors [61].

Unfortunately, deep neural networks are poorly calibrated [106, 74], which cause them to be overconfident in their predictions [39, 106, 74]. One reason for miscalibration of classification models is that the models can overfit the cross-entropy loss easily without overfitting the 0/1 loss (i.e., accuracy) [39, 159]. We propose to add the difference between predicted confidence and accuracy (DCA) as an auxiliary loss for classification model calibration. The DCA term applies a penalty when the cross-entropy loss reduces but the accuracy is plateaued (Figure 6.1).

We evaluate the proposed method across four public medical datasets and four widely used CNN architectures. The results show that our approach reduces calibration error significantly by an average of 65.72% compared to uncalibrated methods (from

0.1006 ECE to 0.0345 ECE), while maintaining the overall accuracy across all the experiments—83.08% and 83.58% for the uncalibrated method and our method, respectively.

6.2 Background

The problem we are addressing is the miscalibration issue of deep neural networks for classification tasks. The confidence associated with a prediction (i.e., probability of being one specific class) should reflect the true correctness likelihood of a model [39]. However, deep neural networks tend to be overconfident in their predictions [106, 74].

6.2.1 Problem Definition

Mathematically, the problem can be defined in the following way. The input $X \in x$ and label $Y \in y = \{1, ..., k\}$ are random variables that follow a joint distribution $\pi(X, Y) = \pi(Y|X)\pi(X)$. Let h be a deep neural network with $h(X) = (\hat{Y}, \hat{P})$, where $\hat{Y}$ is the predicted class label and $\hat{P}$ is the associated confidence. We would like the confidence estimate $\hat{P}$ to be calibrated, which intuitively means that $\hat{P}$ represents a true probability. For instance, given 100 predictions with the average confidence of 0.95, we expect that 95 predictions should be correct. In reality, the average confidence of a deep neural network is often higher than its accuracy [39, 106, 74]. The perfect calibration can be defined as:

$$\mathbb{P}\left(\hat{Y} = Y|\hat{P} = p\right) = p, \forall p \in [0, 1]. \tag{6.1}$$

Difference in expectation between confidence and accuracy (i.e., the calibration error) can be defined as:

$$\mathbb{E}_{\hat{p}}\left[\left|\left(\hat{Y} = Y|\hat{P} = p\right) - p\right|\right]. \tag{6.2}$$

6.2.2 Measurements

Expected Calibration Error (ECE) is a commonly used criterion for measuring neural network calibration error. ECE [99] approximates Equation (6.2) by partitioning predictions into M bins and taking a weighted average of the accuracy/confidence difference for each bin. All the samples need to be grouped into M interval bins according to the predicted probability. Let B_m be the set of indices of samples whose predicted confidence falls into the interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$, $m \in M$. The accuracy of

B_m is

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i), \tag{6.3}$$

where $\hat{y}_i$ and y_i are the predicted and ground-truth label for sample i. The average predicted confidence of bin B_m can be defined as

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i, \tag{6.4}$$

where $\hat{p}_i$ is the confidence of sample i. ECE can be defined with $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right|, \tag{6.5}$$

where n is the number of samples.

Maximum Calibration Error (MCE) [99] is another common criterion for measuring neural network calibration error that partitions predictions into M equally-spaced bins and estimates the worst-case scenario. MCE can be computed as:

$$\mathrm{MCE} = \max_{m \in \{1,\ldots,m\}} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|. \tag{6.6}$$

Comparing with ECE, one of the biggest disadvantages of MCE is that the MCE is very sensitive to the number of bins. By selecting different numbers of bins, users may be able to manipulate the MCE score easily.

6.3 Existing Calibration Methods

In this section, we introduce some existing calibration methods, including temperature scaling [48, 39], entropy regularization [106], MMCE regularization [74], label smoothing [137, 98], and Mixup training [161, 140]. Temperature scaling is a widely used calibration method, which treats model calibration as a post-processing task. All the other methods fix neural network calibration during the classification training stage.

6.3.1 Temperature Scaling

Temperature scaling [48, 39] is a widely-used approach for deep learning model calibration. It fixes the miscalibration issue by dividing the logits by a temperature parameter of T ($T > 0$). The method involves two steps, in general. The first step is to train a classification model. Once the model is trained, the temperature parameter is added to the model and needs to be trained on the validation set while all the other parameters are frozen [39]. After that, the temperature parameter will be used for calibration at the testing time. The calibrated confidence, $\hat{q}_i$, using temperature scaling is

$$\hat{q}_i = \max_k \theta_{SM}(\frac{z_i}{T})^{(k)}, \tag{6.7}$$

where k is the class label ($k = 1, ..., K$), $\theta_{SM}(z_i)$ is the predicted confidence. As $T \rightarrow \infty$, the confidence $\hat{q}_i$ approaches the minimum, which indicates maximum uncertainty.

Temperature scaling is easy to use and performs well. The optimization process of the temperature parameter is not expensive and only needs to be done once. However, as a post-processing approach, temperature scaling does not help with feature learning. In addition, a neural network model should be able to calibrate itself without any post-processing [151].

6.3.2 Trainable Calibration Methods

Trainable calibration methods are proposed to integrate model calibration into classification training. No explicit training round for calibration is needed in such a fashion. One of the earliest trainable approaches is the entropy regularization [106]. The method proposes to use entropy as a regularization term in loss functions for model calibration. The final classification loss can be written as:

$$\text{Loss} = \text{CrossEntropy} + \beta\text{Entropy}, \tag{6.8}$$

where β is a weight scalar. One disadvantage of entropy regularization is that the method is very sensitive to the value of β [74]. Kumar et al. propose to use MMCE replacing entropy for model calibration [74]. MMCE is computed in a reproducing kernel Hilbert space (RKHS) [37]. The completely loss function can be written as:

$$\text{Loss} = \text{CrossEntropy} + \beta(\text{MMCE}^2_m(D))^{\frac{1}{2}}, \tag{6.9}$$

where D denotes a dataset. The performance of MMCE may be limited by imbalance predictions of a neural network. For instance, the number of correct predictions is usually larger than the number of incorrect predictions. Thus, the MMCE term needs to be re-weighted as:

$$\begin{aligned}\mathrm{MMCE}_w^2 = \sum_{c_i=c_j=0} \frac{p_i, p_j, k(p_i, p_j)}{(m-n)^2} + \\ \sum_{c_i=c_j=1} \frac{(1-p_i)(1-p_j)k(p_i, p_j)}{n^2} - \\ 2\sum_{c_i=1, c_j=0} \frac{(1-p_i)p_j k(p_i, p_j)}{(m-n)n},\end{aligned} \tag{6.10}$$

where c is the predicted label, m is the number of correct predictions, n is the batch size, and k is a universal kernel [129].

Label smoothing [137] was proposed to improve classification performance of the Inception architecture [20] Müller et al. demonstrate that label smoothing improves classification performance by calibrating models implicitly [98]. Instead of targeting a hard probability, 1.0, for the correct class, label smoothing tries to predict a softer version of it:

$$y_k^{LS} = y_k(1-\alpha) + \frac{\alpha}{K}, \tag{6.11}$$

where y_k is original targeting probability ($y_k = 1.0$ for the correct class and $y_k = 0.0$ for the rest), K is the number of class labels, α is a hyperparameter that determines the amount of smoothing. Mixup [161] is another method that aims to predict a softer target by randomly mixing training samples. During the training, two samples from different classes are randomly mixed together. Instead of predicting one target label, the network needs to predict the two corresponding labels' probability. The target probabilities equal to the portion of the pixels from each image. Thulasidasan et al. [140] demonstrate that Mixup is also useful for neural network calibration.

6.4 Proposed Method

We propose to add the difference between confidence and accuracy (DCA) as an auxiliary loss term to the cross-entropy loss for classification tasks. DCA is based on expected calibration error by minimizing the difference between the predicted confidence and accuracy directly. Therefore, the proposed method can calibrate neural networks effectively. The proposed method is easy to implement and suitable for any classification tasks. In general, classification loss can be written as follows:

$$\mathrm{Loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + \beta DCA, \tag{6.12}$$

where y_i is the true label and $p(y_i)$ is the predicted confidence (i.e., probability) of the true label. The DCA term can be computed for each mini-batch using the following equation:

$$\text{DCA} = \left| \frac{1}{N} \sum_{i=1}^{N} c_i - \frac{1}{N} \sum_{i=1}^{N} p(\hat{y}_i) \right|, \tag{6.13}$$

where $\hat{y}_i$ is the predicted label; $c_i = 1$, if $\hat{y}_i = y_i$; otherwise, $c_i = 0$. The final loss function can be written as:

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + \beta \left| \frac{1}{N} \sum_{i=1}^{N} c_i - \frac{1}{N} \sum_{i=1}^{N} p(\hat{y}_i) \right|. \tag{6.14}$$

The DCA auxiliary loss fixes the miscalibration issue by penalizing deep learning models when the cross-entropy loss can be reduced, but the accuracy does not change (i.e., when the model is overfitting). The term forces the average predicted confidence to match the accuracy over all training examples without strict constraint on each example, which pushes the network closer to the ideal situation, in which the accuracy reflects the true correctness likelihood of a model. The averaging mechanism of DCA also smooths the predictions that have extremely high or low confidence.

DCA is differentiable in the predicted confidence term but not strictly in the prediction accuracy term due to the argmax step for computing the predicted label. During the training phase, gradients can be backpropagated through the confidence terms but not through the accuracy.

6.5 Experiments

We compare the proposed method with temperature scaling and uncalibrated models (trained with cross-entropy loss without applying any calibration methods) on four medical imaging datasets across four popular CNN networks. The trainable methods are not compared in this work because the literature shows that they have a worse or similar calibration performance with temperature scaling [74, 98, 140]. Thus, it may not be necessary to be compared in this paper explicitly.

The evaluation results show that the proposed method significantly improves model calibration while maintaining the overall classification accuracy. The proposed method reduces calibration error by an average of 65.72% compared to uncalibrated methods (from 0.1006 ECE to 0.0345 ECE) and performs about 20% better than temperature scaling on average.

Table 6.1: Datasets used in this study.

Name	Modality	# of Images	# of Classes
RSNA	Head CT	674257	2
DDSM	Mammography	10480	2
Mendeley V2	Chest X-ray	5856	2
Kather 5000	Histological	5000	8

6.5.1 Experiment Setup

Datasets

Four large, publicly available, medical imaging datasets (RSNA [111], DDSM [46], Mendeley V2 [65], and Kather 5000 [64]) were used in this study for both binary and multi-class classification tasks (Table 6.1).

The RSNA dataset was released for the 2019 RSNA Intracranial Hemorrhage Detection Challenge [111]. We used the training set of the first stage of the data challenge in this study, which contains 674257 CT slices of 17079 patients. The slices were labeled as 7 classes, normal, intracranial hemorrhage, and five sub-classes of intracranial hemorrhage. We used this dataset as a binary classification task (normal/abnormal). The dataset was randomly partitioned into training and testing datasets with a $4 : 1$ ratio on the patient-level by us.

The DDSM dataset contains 2620 well-labeled cases, including 10480 digitized screen-film mammography images [46]. The dataset is almost 20 years old that was initially constructed in 1999. DDSM is the largest publicly available mammography dataset and widely used for developing deep learning models. We chose the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) [80], which is an updated and standardized version of the original DDSM, for our study. We used the training and testing sets provided by the data provider for training and testing respectively. We used the provided training and testing sets in this study.

The Mendeley V2 [65] dataset contains both of the optical coherence tomography (OCT) images of the retina and pediatric chest X-ray images. We used the pediatric chest X-ray images in this study. The dataset includes 4273 pneumonia images and 1583 normal images. We used the provided training and testing sets in this study.

The Kather 5000 [64] dataset contains 5000 histological images of 150×150 pixels. Each image belongs to exactly one of eight tissue categories: tumour epithelium, simple stroma, complex stroma, immune cells, debris, normal mucosal glands, adipose tissue, background (no tissue). All images are RGB, $0.495\mu m$ per pixel, digitized with an

Aperio ScanScope (Aperio/Leica biosystems), magnification 20×. Histological samples are fully anonymized images of formalin-fixed paraffin-embedded human colorectal adenocarcinomas (primary tumors) from the Institute of Pathology, University Medical Center Mannheim, Heidelberg University, Mannheim, Germany). The dataset was randomly partitioned into training and testing datasets with a 4 : 1 ratio by us.

CNN Models

Four transfer learning CNN models were evaluated. More specifically, AlexNet [72], ResNet-50 [44], DenseNet-121 [51], and SqueezeNet 1-1 [53] were used as the feature extractors in the transfer learning models.

We firstly pre-trained these four networks on the ImageNet dataset [20]. Then, the fully connected (FC) layers and the pooling layer before the FC layers were removed from the models. We froze the parameters of the remaining convolutional (Conv) layers of each network and used them as feature extractors. A shallow CNN classifier was trained on top of each feature extractor.

The classifier contained one Conv layer and two FC layers. The Conv layer included convolution, batch normalization [54], leaky ReLU [155], and max pooling [72]. Max pooling had a 2×2 receptive field with stride 1. Dropout [131] with a rate of 0.5 was applied to the FC layers. Weighted cross-entropy loss was used in the training. Adam optimizer [68] with a learning rate of 0.0001 was used as the optimizer.

For the same architecture, all the hyper-parameters were maintained the same among different datasets, except batch sizes. The batch size of AlexNet on DDSM, Mendeley V2, and Kather 5000 datasets was set as 64, and for RSNA was 512. For the rest of the architectures, the batch sizes were set as half of the AlexNet with the corresponding datasets.

6.5.2 Calibration Results

Table 6.2 shows the expected calibration error (ECE) and the accuracy of the uncalibrated models (Unca.), temperature scaling (Temp.), and the proposed method (DCA). Each model was trained for two times. The average value is shown in the table.

The table shows that our method is consistently better than the uncalibrated method on model calibration, which reduces the ECE by 65.71% on average (from 0.1006 to 0.0345). Temperature scaling has the second smallest average ECE (0.0427). However, it is still 23.77% worse than the proposed method. On average, the uncali-

Table 6.2: Expected Calibration Error (ECE) for Each Model

Dataset	**Model**	**ECE** (smaller is better)			**Accuracy**[1] (larger is better)	
		Unca.	**Temp.**	**DCA**	**Unca.**	**DCA**
RSNA	AlexNet	**0.0113**	0.0239	0.0120	0.8376	**0.8488**
	ResNet	0.0276	0.0231	**0.0122**	0.8569	**0.8762**
	DenseNet	0.0102	0.0814	**0.0077**	0.8502	**0.8543**
	SqueezeNet	0.0253	0.0317	**0.0097**	0.8671	**0.8841**
DDSM	AlexNet	0.2164	0.0658	**0.0591**	**0.6766**	0.6291
	ResNet	0.1844	**0.0307**	0.0798	**0.7195**	0.6987
	DenseNet	0.1798	**0.0337**	0.0754	0.7076	**0.7106**
	SqueezeNet	0.2173	**0.0458**	0.0805	**0.6853**	0.6771
Mendeley	AlexNet	0.1693	0.0396	**0.0273**	0.8585	**0.8785**
	ResNet	0.1475	0.0475	**0.0291**	0.8520	**0.8767**
	DenseNet	0.1136	0.0746	**0.0285**	0.8331	**0.8796**
	SqueezeNet	0.1871	0.0468	**0.0252**	0.8742	**0.8750**
Kather	AlexNet	0.0279	0.0344	**0.0243**	**0.9062**	0.9052
	ResNet	**0.0248**	0.0318	0.0304	**0.9355**	0.9229
	DenseNet	0.0302	0.0286	**0.0237**	0.9385	**0.9410**
	SqueezeNet	0.0372	0.0439	**0.0269**	0.8932	**0.9038**
Average		0.1006	0.0427	**0.0345**	0.8308	**0.8351**

[1] The temperature scaling method has the same accuracy as the uncalibrated models.

brated method and temperature scaling have an 83.08% accuracy, while the proposed method has an 83.51% accuracy. The proposed method increases the accuracy of 11 out of 16 tests.

It is worth noting that temperature scaling increases calibration error of 3 out of 4 models on the Kather 5000 dataset, while the proposed method is still able to reduce the calibration error of most cases on the same dataset. The Kather 5000 dataset is a relatively simple and large dataset for its task. The dataset is considered as the MNIST of histology images. It is speculated to have a sufficient amount of training data to train a model end-to-end, with a smaller overfitting effect (i.e., miscalibration). In such a case, since the temperature parameter (T) of temperature scaling is learned on only the validation set, it may actually hurt the calibration. However, the proposed method jointly optimizes the accuracy and modal calibration simultaneously, and it can still reduce the calibration error.

Table 6.3 shows the maximum calibration error (MCE) of the compared models trained using the RSNA, DDSM, and Mendeley datasets. Though there is no clear winner on MCE, the proposed model decreases the average MCE by about 3%, while temperature scaling increases the MCE slightly. Table 6.4 shows the MCE results

Table 6.3: Maximum Calibration Error (MCE) for Binary Classification Tasks

Dataset	**Model**	**MCE** (smaller is better)		
		Unca.	**Temp.**	**DCA**
RSNA	AlexNet	0.0291	0.0366	**0.0230**
	ResNet	0.0484	**0.0325**	0.0399
	DenseNet	0.0335	0.2233	**0.0142**
	SqueezeNet	0.0430	0.0663	**0.0270**
DDSM	AlexNet	0.2527	**0.1545**	0.1800
	ResNet	0.2897	0.1171	**0.1078**
	DenseNet	0.2403	**0.0941**	0.0959
	SqueezeNet	0.332	0.1631	**0.1586**
Mendeley	AlexNet	0.2454	0.4305	**0.1225**
	ResNet	0.2321	**0.1769**	0.297
	DenseNet	0.2653	**0.2477**	0.4898
	SqueezeNet	0.2521	**0.2451**	0.2507
Average		0.2812	0.2817	**0.2737**

Table 6.4: Maximum Calibration Error for Multi-Class Classification Tasks

Dataset	**Model**	**MCE** (smaller is better)		
		Unca.	**Temp.**	**DCA**
Kather 5000	AlexNet	**0.2570**	0.2974	0.7371
	ResNet	0.3072	0.6379	**0.2513**
	DenseNet	**0.2577**	0.8015	0.4268
	SqueezeNet	**0.2566**	0.3237	0.7371
Average		**0.2696**	0.5151	0.5381

on the Kather 5000 dataset. Both temperature scaling and the proposed method increase the MCE quite well. According to Guo et al., MCE is very sensitive to the number of bins since it measures the worst cases across all bins, making it an improper metric for small test sets [39]. On average, temperature scaling has an MCE of 0.3401 across all the evaluated models, the proposed method has an MCE of 0.3398, and the uncalibrated method has an MCE of 0.2766.

6.5.3 Model Representation Learning

As a post-processing method, temperature scaling fixes model miscalibration at the output-level, which does not change the learned representation. However, the proposed method integrates calibration into the network training phase, which may help models

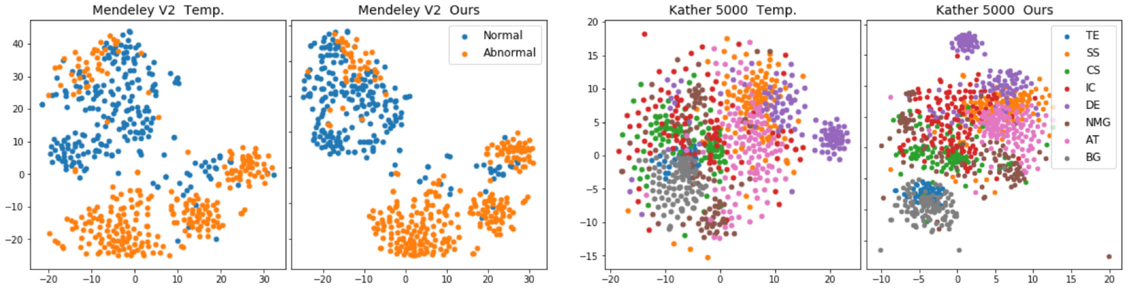


Figure 6.2: The t-SNE plots of the representations learned using temperature scaling and the proposed method on the Mendeley V2 (left) and Kather 5000 (right) datasets. The samples of temperature scaling are spreading in the feature space (left in each subplot). The samples of the proposed method are densely packed for the same class (right in each subplot).

learn a better representation. In this section, we firstly use t-SNE [88] plots to visualize features extracted by temperature scaling and DCA on two datasets. Then, we compare the recovered probability distribution of the two methods on a toy dataset.

Figure 6.2 shows the t-SNE plots of features extracted by temperature scaling and the proposed method on the Mendeley V2 and Kather 5000 datasets. Each dot represents one data sample in a 2D feature space. To generate these plots, We first extract the high-dimensional feature maps using an AlexNet model trained with either temperature scaling or DCA. We then feed the feature maps to t-SNE which projects the high-dimensional feature maps to 2D feature space. Ideally, the samples from the same class should be close to each other; the samples from different classes should be far from each other.

The plots reveal that the samples of temperature scaling (left in each subplot) are spreading in the feature space regardless of class labels, while the samples of the same class are densely packed in our method (right in each subplot). Especially for the Kather 5000 dataset, DCA (the rightmost figure) successfully separated the samples of TE (blue) and GB (gray) classes from the rest. But, the same classes of temperature scaling are mixed with others.

Figure 6.3 shows the probability distribution recovered by the uncalibrated method (left), temperature scaling (middle), and the proposed method (right) on a toy dataset. For this experiment, we train three simple networks using the uncalibrated method, temperature scaling, and the proposed method, separately. The networks share the same two-layer architecture that takes a one-dimensional input for a binary classification task. We randomly sample a dataset between -2 and 2, and randomly label each sample with either 0 or 1. The curved line in each figure shows the recovered

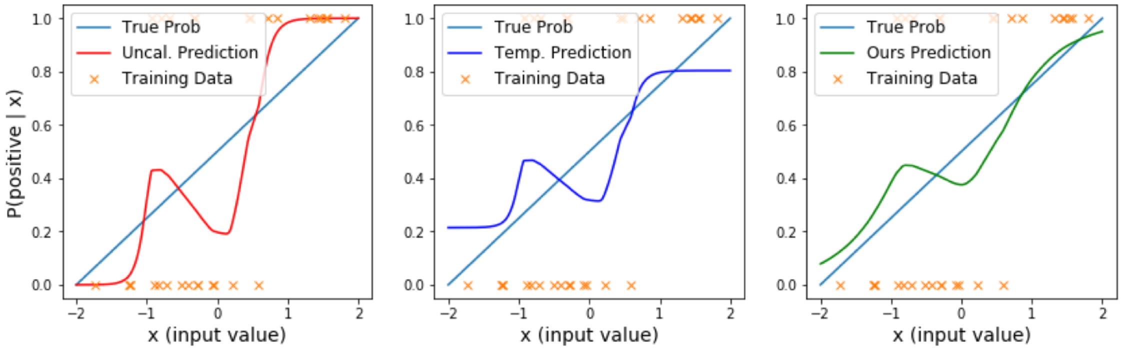


Figure 6.3: The figure shows the probability distribution that was recovered by the uncalibrated method (left), temperature scaling (middle), and the proposed method (right). The recovered distribution of the uncalibrated model (left) is far from the ground-truth with many overconfident predictions. Temperature scaling (middle) reduces the predicted confidence of the uncalibrated model, but the recovered distribution is still far from the ground-truth. Our method (right) can better recover the true probability.

probability distribution, while the light blue line in each figure shows the ground-truth distribution.

The figures reveal that the uncalibrated model (left) has many high-probability predictions. For instance, when $-2 < x < -1$, the model made many negative predictions with high-probability; when $1 < x < 2$, the model made many positive predictions with high-probability. The majority of these predicted probabilities are far from the true probabilities, which indicates the model is overconfident in its predictions and does not capture the true probability distribution well. The temperature scaling method (middle) can relax those extreme predictions by pushing the predicted probabilities close to 0.5. However, the recovered probability distribution is still quite far from the diagonal line. The proposed method (right) can recover the trend of the ground-truth distribution, and most of the predictions are close to the diagonal line. This experiment shows the models trained with DCA may have a strong ability to recover the true probability distribution more accurately.

6.5.4 Hyperparameter Effects

One drawback of the proposed method is that the weight scalar β needs to be selected for each model. In this section, we show the testing result of the proposed method with different weights ($\beta = [1, 5, 10, 15, 20, 25]$). Figure 6.4 shows the train/test accuracy and loss on the Mendeley V2 dataset using AlexNet architecture with four different β

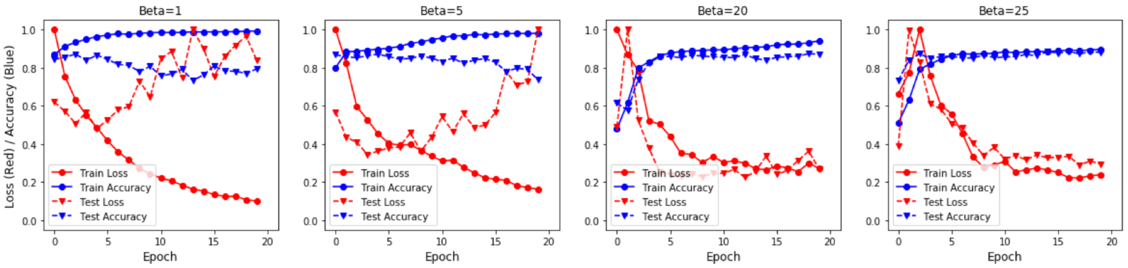


Figure 6.4: Train/Test accuracy and loss of Mendeley V2 dataset using AlexNet architecture with four different β values.

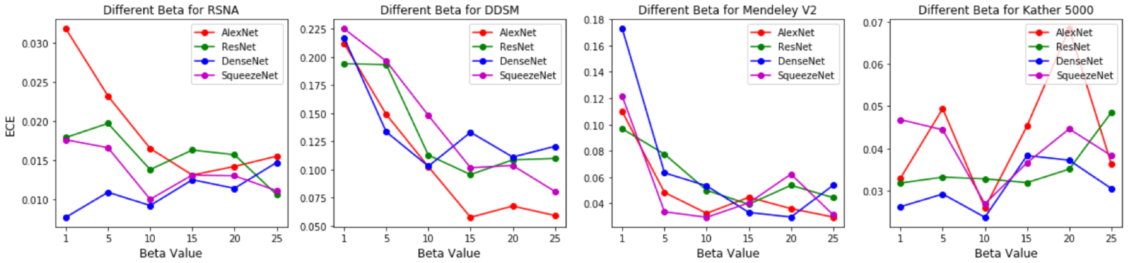


Figure 6.5: Expected calibration error (ECE) of each dataset with different β values.

values. From the results, we can see that a smaller value such as 1 or 5 usually will not be a good choice since they put a smaller penalty to the model when the cross-entropy loss is overfitted. Among our experiments, most of the best results appeared using a weight between 10 and 25.

Figure 6.5 shows the ECE results of all of the evaluated models with different β values. The figure reveals that the ECE result is not very sensitive to the β value when $\beta \geq 10$, except for the Kather 5000 dataset. In our experiences, we use β values between 10 and 15 for most of the tasks.

6.6 Conclusion

We proposed a novel approach to neural network calibration that maintains classification accuracy while significantly reducing model calibration error. We evaluated our approach across various architectures and datasets. The results show that our approach reduces calibration error significantly and comes closer to recovering the true probability than other approaches. The proposed method can be easily integrated into any classification tasks as an auxiliary loss term, thus not requiring an explicit training round for calibration. We believe this simple, fast, and straightforward method can serve as a strong baseline for future researchers.

Chapter 7

Discussion

The research in this dissertation investigated the effectiveness of utilizing state-of-the-art deep learning techniques on medical imaging analysis. The major challenges for such tasks include 1) lack of annotated data in the medical imaging domain, 2) how to using domain knowledge in deep learning models, and 3) how much we can trust in deep learning models. We proposed multiple novel deep learning models to address the challenges mentioned above. Each model has its own advantages towards using deep learning models in physicians' daily clinical practices.

In Chapter 3, we introduced a multi-modal classification framework, which evaluates 2D and pseudo-3D data simultaneously. The proposed framework is inspired by the daily clinical practice of specialized breast radiologists. Two imaging modalities, mammogram and digital breast tomosynthesis (DBT), are widely used in clinical practice when diagnosing breast cancer. Doctors usually need to read both modalities before making a decision. However, almost all of the existing deep learning models, before this work, are focusing on using only one modality. Our approach successfully uses the two modalities simultaneously. The architecture design was motivated by domain experts and domain knowledge. Our experiments show that, similar to human doctors, the performance of deep learning models can be improved when using both imaging modalities. Other than using a private dataset, one limitation of this work is the method we used to convert DBT to a fixed-size representation is not trainable. We think a learning-based method can further improve the model performance.

In Chapter 4, we proposed a weakly-supervised breast cancer detection network. Conventionally, object detection networks are trained supervised with bounding boxes. However, such fine-grained annotations usually do not exist in the clinical relevance dataset due to the high annotation cost. We proposed a class activation mapping (CAM) based method to train a lesion detection network using only image-level labels.

CAM is a widely used building block for weakly supervised object detection networks. However, comparing with many other imaging domains, the ROIs in breast imaging are very small, and the ratio of image size to ROI size is often much higher. Apply CAM naively may not give us a precise localization result. Instead of using CAM naively, we use the CAM as the starting point and use a self-training strategy to refine the predicted step-by-step. Our results show that the proposed model significantly improves performance compared to other methods trained similarly.

In Chapter 5, we presented a general image feature learning framework that learns the feature representations from a large set of images from the same domain through a text and image matching network. Manual annotation is costly in the medical domain. Deep learning researchers are always suffering from the lack of labeled data issues. However, plenty of text data exists in a medical record system that provides rich information about medical images, but it is hard to use to train a deep learning model directly. The proposed framework learns image feature representation through an image and text matching network. The learned feature representation can be used to build various downstream applications. The downstream applications can be trained with tiny annotated datasets. Our result shows that we are able to significantly reduce the need for manually labeled data using our strategy; in some cases, the proposed method reduces the need for manually annotated training data by 99%.

In Chapter 6, we explored a problem related to safe AI in the medical domain. How much we can trust a deep learning model is a common question that has been asked by medical experts. Deep learning models report exciting performance on many tasks. However, uncertainty quantification is often ignored when evaluating these models. Unfortunately, modern deep learning networks tend to be overconfident in their predictions, which could be problematic in an automatic decision-making system, especially for the medical field. The current state-of-the-art method, temperature scaling, for deep learning calibration solves the problem by adding a post-processing step, which does not affect the representation ability of a model. We proposed a trainable calibration method, which integrated deep learning mode calibration into the training stage. Our evaluation result shows that the proposed method is not only able to improve deep learning model calibration, but may also improve the representation ability of a deep learning model.

In this dissertation, we address three different challenges of applying deep learning in medical imaging analysis in four different chapters. There are many options to extend our work. For instance, in Chapter 3, we converted DBTs with various lengths to a fixed size representation using RankSVM. The method works well in this task. However, one disadvantage is that no task-specific learnable parameters are involved in this process. We think that by including some trainable parameters into the conversion process that can make the converted data format more task-specific, which may further improve the model performance. In Chapter 4 and Chapter 5, we introduced two different methods to deal with the limited training data issue in the medical imaging domain. Especially in Chapter 5, we proposed an image feature learning method to use the weak supervisions of existing text documents. Though this method significantly reduced the need for manual annotations, the method still relies on some forms of annotations that can be got easily. Another direction to solve this problem will be learning the feature representation using unsupervised methods. In such a way, we can completely get rid of any forms of supervision. We proposed a neural network calibration method in Chapter 6, which performs better than the state-of-the-art calibration method. It helps a prediction model to get a better estimate of its prediction confidence. However, the proposed method requires a pre-selected weight for the regularization term. We may need to use different weights for different models, which becomes a bottleneck of the proposed method. In the future, we hope to use a neural architecture search method to select an optimal weight for each model automatically.

We think there are still plenty of works that need to be done before using deep learning in clinical practice. However, we believe the future will be bright. We hope that our work will inspire other researchers to continue investigating how to apply deep learning in the medical imaging domain.

Bibliography

[1] Artificial intelligence and machine learning-based medical devices: A products liability perspective. 4

[2] Breast cancer screening guidelines. https://www.cancer.org/health-care-professionals/american-cancer-society-prevention-early-detection-guidelines/breast-cancer-screening-guidelines.html. 14

[3] Mammograms. https://www.cancer.gov/types/breast/mammograms-fact-sheet. 16

[4] Hologic receives fda approval for first 3-d digital mammography (breast tomosynthesis) system. *Hologic website.*, 2011. 16

[5] The digital mammography dream challenge. https://www.synapse.org/#!Synapse:syn4224222/wiki/401743, Nov 2016. 17

[6] Medical imaging. https://www.who.int/diagnostic_imaging/en, Feb 2017. 1

[7] The cancer imaging archive. https://www.cancerimagingarchive.net/, Oct 2020. 4, 37

[8] Wenjia Bai et al. Semi-supervised learning for network-based cardiac MR image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017. 39

[9] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. 28

[10] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019. 39

[11] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. Dynamic image networks for action recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 3034–3042. IEEE, 2016. 18

[12] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 2846–2854. IEEE, 2016. 28

[13] Aaron F Bobick and James W Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):257–267, 2001. 18

[14] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018. 14

[15] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 6299–6308. IEEE, 2017. 17

[16] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019. 39

[17] Veronika Cheplygina, Isabel Pino Pena, Jesper Holst Pedersen, David A Lynch, Lauge Sørensen, and Marleen de Bruijne. Transfer learning for multicenter classification of chronic obstructive pulmonary disease. *IEEE Journal of Biomedical And Health Informatics*, 22(5):1486–1496, 2017. 39

[18] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, volume 1, pages 539–546. IEEE, 2005. 12

[19] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 17

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the Conference on*

Computer Vision and Pattern Recognition, pages 248–255. Ieee, 2009. 7, 10, 20, 58, 61

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 40, 41

[22] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211, 2007. 1

[23] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 642–651. IEEE, 2017. 28, 33, 39

[24] Olivier Ecabert et al. Automatic model-based segmentation of the heart in ct images. *IEEE transactions on medical imaging*, 27(9):1189–1201, 2008. 1

[25] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017. 2, 7, 27, 37, 54

[26] Thorsten Falk et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019. 37

[27] US FDA. Proposed regulatory framework for modifications to artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd). FDA, 2019. 4

[28] Basura Fernando, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars. Rank pooling for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):773–787, 2016. 19

[29] Basura Fernando, Efstratios Gavves, Jose M Oramas, Amir Ghodrati, and Tinne Tuytelaars. Modeling video evolution for action recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 5378–5387. IEEE, 2015. 18

[30] Alejandro F Frangi, Wiro J Niessen, and Max A Viergever. Three-dimensional modeling for functional analysis of cardiac images, a review. *IEEE transactions on medical imaging*, 20(1):2–5, 2001. 1

[31] Giovanni B Frisoni, Nick C Fox, Clifford R Jack, Philip Scheltens, and Paul M Thompson. The clinical use of structural mri in alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010. 1

[32] Maryellen L Giger, Heang-Ping Chan, and John Boone. Anniversary paper: history and status of cad and quantitative image analysis: the role of medical physics and aapm. *Medical physics*, 35(12):5799–5820, 2008. 1

[33] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 28, 40

[34] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 580–587. IEEE, 2014. 40

[35] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 9

[36] W Gray, CM Rumack, SR Wilson, and JW Charboneau. *Diagnostic ultrasound*. New York: Mosby, 1998. 1

[37] Arthur Gretton. Introduction to rkhs, and some simple kernel algorithms. *Adv Top. Mach. Learn. Lecture Conducted from University College London*, 2013. 57

[38] Lin Gu, Yinqiang Zheng, Ryoma Bise, Imari Sato, Nobuaki Imanishi, and Sadakazu Aiso. Semi-supervised learning for biomedical image segmentation via forest oriented super pixels (voxels). In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 702–710. Springer, 2017. 39

[39] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 1321–1330. JMLR, 2017. 4, 54, 55, 56, 57, 63

[40] Philipp Hacker, Ralf Krestel, Stefan Grundmann, and Felix Naumann. Explainable ai under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law*, pages 1–25, 2020. 4

[41] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006. 12

[42] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65, 2019. 37

[43] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 6546–6555. IEEE, 2018. 17

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016. 7, 22, 40, 43, 61

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision*, pages 630–645. Springer, 2016. 2

[46] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and W Philip Kegelmeyer. The digital database for screening mammography. In *Proceedings of the 5th international workshop on digital mammography*, pages 212–218. Medical Physics Publishing, 2000. 60

[47] P Henrot, A Leroux, C Barlier, and P Génin. Breast microcalcifications: the lesions in anatomical pathology. *Diagnostic and interventional imaging*, 95(2):141–152, 2014. 1, 14

[48] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 56

[49] Minh Hoai and Andrew Zisserman. Improving human action recognition using score distribution and ranking. In *Asian conference on computer vision*, pages 3–20. Springer, 2014. 18

[50] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 40

[51] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 4700–4708. IEEE, 2017. 22, 61

[52] Sangheum Hwang and Hyo-Eun Kim. Self-transfer learning for weakly supervised lesion localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 239–246. Springer, 2016. 12, 33, 34

[53] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 22, 61

[54] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 20, 61

[55] Jeremy Irvin et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019. 38

[56] Mohammadhassan Izadyyazdanabadi et al. Weakly-supervised learning-based feature localization for confocal laser endomicroscopy glioma images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–308, 2018. 39

[57] Gareth James. Majority vote classifiers: theory and applications. *Research thesis*, 1998. 21

[58] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018. 28

[59] Zhipeng Jia, Xingyi Huang, I Eric, Chao Chang, and Yan Xu. Constrained deep weak supervision for histopathology image segmentation. *IEEE Transactions on Medical Imaging*, 36(11):2376–2388, 2017. 39

[60] Bo Jiang, Ziyan Zhang, Doudou Lin, Jin Tang, and Bin Luo. Semi-supervised learning with graph learning-convolutional networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 11313–11320. IEEE, 2019. 39

[61] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2011. 54

[62] Alistair EW Johnson et al. Mimic-cxr: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 1(2), 2019. 37

[63] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems*, pages 1889–1897, 2014. 40

[64] Jakob Nikolas Kather et al. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6:27988, 2016. 37, 60

[65] DK Kermany and M Goldbaum. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley Data*, 2, 2018. 37, 60

[66] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 6092–6101. IEEE, 2019. 39

[67] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 12456–12465. IEEE, 2019. 28

[68] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 20, 44, 61

[69] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 40

[70] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *Proceedings of the International conference on machine learning deep learning workshop*, volume 2, 2015. 13

[71] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. In *Advances in Neural Information Processing Systems*, pages 4225–4235, 2017. 39

[72] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 7, 9, 17, 20, 22, 61

[73] Elizabeth A Krupinski. Current perspectives in medical image perception. *Attention, Perception, & Psychophysics*, 72(5):1205–1217, 2010. 1

[74] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the International Conference on Machine Learning*, pages 2810–2819, 2018. 4, 54, 55, 56, 57, 59

[75] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*, pages 33–40. IEEE, 2016. 13

[76] C LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 9

[77] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015. 2

[78] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 41

[79] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pages 201–216, 2018. 40

[80] R Sawyer Lee, Francisco Gimenez, Assaf Hoogi, and Daniel Rubin. Curated breast imaging subset of ddsm. *The Cancer Imaging Archive*, 2016. 60

[81] Gongbo Liang, Sajjad Fouladvand, Jie Zhang, Michael A Brooks, Nathan Jacobs, and Jin Chen. Ganai: Standardizing ct images using generative adversarial network with alternative improvement. In *International Conference on Healthcare Informatics*, pages 1–11. IEEE, 2019. 37, 54

[82] Gongbo Liang, Xiaoqin Wang, Yu Zhang, and Nathan Jacobs. Weakly-supervised self-training for breast cancer localization. In *Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, pages 1124–1127. IEEE, 2020. 12

[83] Gongbo Liang, Xiaoqin Wang, Yu Zhang, Xin Xing, Hunter Blanton, Tawfiq Salem, and Nathan Jacobs. Joint 2d-3d breast cancer classification. In *International Conference on Bioinformatics and Biomedicine*, pages 692–696. IEEE, 2019. 54

[84] Gongbo Liang, Yu Zhang, Xiaoqin Wang, and Nathan Jacobs. Improved trainable calibration method for neural networks on medical imaging classification. In *British Machine Vision Conference*. BMVC, 2020. 27

[85] Geert Litjens et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. 4, 37, 38

[86] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 10921–10930. IEEE/CVF, 2020. 40

[87] Zhen Ma, João Manuel RS Tavares, Renato Natal Jorge, and T Mascarenhas. A review of algorithms for medical image segmentation and their applications to the female pelvic cavity. *Computer Methods in Biomechanics and Biomedical Engineering*, 13(2):235–246, 2010. 1

[88] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 64

[89] Saket Maheshwary and Hemant Misra. Matching resumes to jobs via deep siamese network. In *Companion Proceedings of the The Web Conference 2018*, pages 87–88, 2018. 13

[90] Jeanne S Mandelblatt et al. Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. *Annals of internal medicine*, 151(10):738–747, 2009. 1

[91] Maciej A Mazurowski, Piotr A Habas, Jacek M Zurada, Joseph Y Lo, Jay A Baker, and Georgia D Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, 21(2-3):427–436, 2008. 4

[92] Kayla Mendel, Hui Li, Deepa Sheth, and Maryellen Giger. Transfer learning from convolutional neural networks for computer-aided diagnosis: a comparison of digital breast tomosynthesis and full-field digital mammography. *Academic radiology*, 26(6):735–743, 2019. 14, 17

[93] Diana L Miglioretti et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. *JAMA pediatrics*, 167(8):700–707, 2013. 1

[94] Diana L Miglioretti et al. Digital breast tomosynthesis: radiologist learning curve. *Radiology*, 291(1):34–42, 2019. 16

[95] Radu Paul Mihail, Gongbo Liang, and Nathan Jacobs. Automatic hand skeletal shape estimation from radiographs. *IEEE transactions on nanobioscience*, 18(3):296–305, 2019. 27, 54

[96] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations*, 2013. 40

[97] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013. 40

[98] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Proceedings of the Advances in Neural Information Processing Systems Systems*, pages 4694–4703, 2019. 56, 58, 59

[99] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 55, 56

[100] Emilio Soria Olivas. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques*. IGI Global, 2009. 10

[101] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 685–694. IEEE, 2015. 28, 33

[102] Niall O'Mahony et al. Deep learning vs. traditional computer vision. In *Science and Information Conference*, pages 128–144. Springer, 2019. 2

[103] Adam Paszkeand et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035. 2019. 22, 43

[104] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188, 2018. 38

[105] Heather R Peppard, Brandi E Nicholson, Carrie M Rochman, Judith K Merchant, Ray C Mayo III, and Jennifer A Harvey. Digital breast tomosynthesis in the diagnostic setting: indications and clinical applications. *Radiographics*, 35(4):975–990, 2015. 16

[106] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. 4, 54, 55, 56, 57

[107] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 779–788. IEEE, 2016. 28

[108] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 17, 40

[109] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):4165, 2018. 7, 14, 17, 28, 54

[110] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 7, 54

[111] RSNA. Rsna intracranial hemorrhage detection. https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/overview, 2019. 60

[112] Andrew Rubin and Andrew Rubin. History of medical imaging - a brief overview. https://www.flushinghospital.org/newsletter/history-of-medical-imaging-a-brief-overview/, Apr 2017. 1

[113] Olga Russakovsky et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 28

[114] Michael S Ryoo, Brandon Rothrock, and Larry Matthies. Pooled motion features for first-person videos. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 896–904. IEEE, 2015. 18

[115] Berkman Sahiner et al. Deep learning in medical imaging and radiation therapy. *Medical physics*, 46(1):e1–e36, 2019. 27, 54

[116] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 5814–5824. IEEE, 2019. 40

[117] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015. 7

[118] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 815–823. IEEE, 2015. 12, 13

[119] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 11, 28, 39

[120] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009. 39

[121] Li Shen. End-to-end training for whole image breast cancer diagnosis using an all convolutional design. *arXiv preprint arXiv:1708.09427*, 2017. 17

[122] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019. 14

[123] David Silver et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016. 2

[124] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 9, 17

[125] Per Skaane et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology*, 267(1):47–56, 2013. 16

[126] Rebecca Smith-Bindman et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. *Jama*, 307(22):2400–2409, 2012. 1

[127] Rebecca Smith-Bindman et al. Ultrasonography versus computed tomography for suspected nephrolithiasis. *New England Journal of Medicine*, 371(12):1100–1110, 2014. 1

[128] Rebecca Smith-Bindman et al. Trends in use of medical imaging in us health care systems and in ontario, canada, 2000-2016. *Jama*, 322(9):843–856, 2019. 1

[129] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007. 58

[130] Shelly Soffer, Avi Ben-Cohen, Orit Shimon, Michal Marianne Amitai, Hayit Greenspan, and Eyal Klang. Convolutional neural networks for radiologic images: a radiologist's guide. *Radiology*, 290(3):590–606, 2019. 7

[131] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 20, 61

[132] Gary M Strauss, Ray E Gleason, and David J Sugarbaker. Chest x-ray screening improves outcome in lung cancer: a reappraisal of randomized trials on lung cancer screening. *Chest*, 107(6):270S–279S, 1995. 1

[133] Yuanyuan Su et al. A deep learning view of the census of galaxy clusters in illustristng. *Monthly Notices of the Royal Astronomical Society*, 498(4):5620–5628, 2020. 7

[134] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019. 41

[135] Kenji Suzuki. Overview of deep learning in medical imaging. *Radiological physics and technology*, 10(3):257–273, 2017. 54

[136] Stephen J Swensen et al. Lung nodule enhancement at ct: multicenter study. *Radiology*, 214(1):73–80, 2000. 1

[137] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 2818–2826. IEEE, 2016. 9, 56, 58

[138] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1701–1708. IEEE, 2014. 12

[139] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, page 101693, 2020. 39

[140] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Proceedings of the Advances in Neural Information Processing Systems Systems*, pages 13888–13899, 2019. 56, 58, 59

[141] Bram Van Ginneken, Alejandro F Frangi, Joes J Staal, Bart M ter Haar Romeny, and Max A Viergever. Active shape model segmentation with optimal features. *IEEE transactions on medical imaging*, 21(8):924–933, 2002. 1

[142] Armine Vardazaryan, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Weakly-supervised learning for tool localization in laparoscopic videos. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 169–179. Springer, 2018. 12, 33, 34

[143] Ashish Vaswani et al. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 28, 41

[144] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 2199–2208. IEEE, 2019. 39

[145] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018. 40

[146] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1508–1517. IEEE, 2020. 40

[147] Xiaoqin Wang, Gongbo Liang, Yu Zhang, Hunter Blanton, Zachary Bessinger, and Nathan Jacobs. Inconsistent performance of deep learning models on mammogram classification. *Journal of the American College of Radiology*, 2020. 4, 28, 37

[148] Jürgen Weese and Cristian Lorenz. Four challenges in medical image analysis from an industrial perspective. *Medical Image Analysis*, 33:44–49, 2016. 37, 38

[149] Jônatas Wehrmann and Rodrigo C Barros. Bidirectional retrieval made simple. In *IEEE Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 7718–7726. IEEE, 2018. 40

[150] Melissa Whitworth, Leanne Bricker, and Clare Mullan. Ultrasound for fetal assessment in early pregnancy. *Cochrane database of systematic reviews*, (7), 2015. 1

[151] David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. In *Advances in Neural Information Processing Systems*, pages 12236–12246, 2019. 57

[152] Martin J Willemink et al. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020. 4, 37

[153] Thomas Wolf et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv:1910.03771*, 2019. 43

[154] Xin Xing et al. Dynamic image for 3d mri image alzheimer's disease classification. In *Proceedings of the European Conference on Computer Vision Workshops*, 2020. 27, 54

[155] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 20, 61

[156] Qingsong Yang et al. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, 37(6):1348–1357, 2018. 7, 27, 54

[157] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 93–102. IEEE, 2019. 39

[158] Ying Yu, Min Li, Liangliang Liu, Yaohang Li, and Jianxin Wang. Clinical big data and deep learning: Applications, challenges, and future outlooks. *Big Data Mining and Analytics*, 2(4):288–305, 2019. 4

[159] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016. 54

[160] Dingwen Zhang, Deyu Meng, and Junwei Han. Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):865–878, 2016. 39

[161] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations*, 2018. 56, 58

[162] Xiaofei Zhang, Yi Zhang, Erik Y Han, Nathan Jacobs, Xiaoqin Han, and Jinze Liu. Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks. *IEEE transactions on nanobioscience*, 17(3):237–242, 2018. 14, 17, 18, 22

[163] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European Conference on Computer Vision*, pages 597–613, 2018. 28

[164] Yu Zhang, Gongbo Liang, Tawfiq Salem, and Nathan Jacobs. Defense-pointnet: Protecting pointnet against adversarial attacks. In *2019 IEEE International Conference on Big Data*, pages 5654–5660. IEEE, 2019. 37

[165] Yu Zhang, Gongbo Liang, Yuanyuan Su, and Nathan Jacobs. Multi-branch attention networks for classifying galaxy clusters. In *Proceeding of International Conference on Pattern Recognition*. IEEE, 2021. 7

[166] Yu Zhang, Xiaoqin Wang, Hunter Blanton, Gongbo Liang, Xin Xing, and Nathan Jacobs. 2d convolutional neural networks for 3d digital breast tomosynthesis classification. In *International Conference on Bioinformatics and Biomedicine*, pages 1013–1017. IEEE, 2019. 14, 27, 54

[167] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 2921–2929. IEEE, 2016. 11, 28, 31, 39, 48

[168] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018. 39

[169] Wentao Zhu, Chaochun Liu, Wei Fan, and Xiaohui Xie. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In *Winter Conference on Applications of Computer Vision*, pages 673–681. IEEE, 2018. 28

[170] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005. 39

Vita

Gongbo “Tony" Liang

Education

† *Indicates expected*

08/2016–12/2020†	Ph.D., Computer Science, University of Kentucky, USA
08/2013–05/2016	M.S., Computer Science, Western Kentucky University, USA
08/2010–01/2013	M.A., Folk Studies, Western Kentucky University, USA
09/2004–07/2008	B.A, Video Game Design, Northeastern University, China

Appointments

11/2020 – Present	Assistant Professor, Computer Science, Eastern Kentucky University
08/2016 – 05/2020	Graduate Assistant, Computer Science, University of Kentucky
08/2013 – 05/2016	Graduate Assistant, Computer Science, Western Kentucky University
09/2012 – 12/2012	Intern, Folklife and Folk Music, Houston Arts Alliance
09/2008 – 08/2009	Lecture, Shenyang Technicians College, Shenyang, China
07/2007 – 09/2008	Assistant 3D Model Designer, Hima Technology Co., Shenyang, China

Publications

Refereed Journals

[1] Y. Su, Y. Zhang, **Gongbo Liang**, J. ZuHone, D. Barnes, N. Jacobs, M. Ntampaka et al. “A deep learning view of the census of galaxy clusters in IllustrisTNG."*Monthly Notices of the Royal Astronomical Society.* (2020). Oxford University Press. Impact Factor: 5.356 (2019). DOI: *Available soon.*

[2] T. Hammond, X. Xing, D. Ma1, K. Nho, F. Elahi, D. Ziegler, **Gongbo Liang**, Q. Cheng, N. Jacobs, P. Crane, ADNI, A. Lin. “β -Amyloid and Tau Drive Early Alzheimer’s Disease Decline While Glucose Hypometabolism Drives Late Decline.” In *Communications Biolog* 3, no. 1 (2020): 1-13. Nature Research. DOI: 10.1038/s42003-020-1079-x

[3] X. Wang, **Gongbo Liang**, Y. Zhang, H. Blanton, Z. Bessinger, and N. Jacobs. “Inconsistent Performance of Deep Learning Models on Mammogram Classification.” In *Journal of the American College of Radiology* 17, no. 6 (2020): 796-803. Elsevier. Impact Factor: 4.268. DOI: 10.1016/j.jacr.2020.01.006

[4] R. Mihail, **Gongbo Liang**, N. Jacobs. “Automatic Hand Skeletal Shape Estimation from Radiographs.” In *IEEE Transactions on NanoBioscience* 18, no 3, pp. 296-305, IEEE, 2019. Impact Factor: 2.791. DOI: 10.1109/TNB.2019.2911026

Refereed Conferences

[1] Y. Zhang, **Gongbo Liang**, Y. Su, and N. Jacobs. “Parametric Attention for Sparse Image Classification." In *25th International Conference on Pattern Recognition (ICPR)*, 2021. Milan, Italy. DOI: Available soon

[2] **Gongbo Liang**, Y. Zhang, X. Wang, and N. Jacobs . “Improved Trainable Calibration Method for Neural Networks on Medical Imaging Classification." In *2020 31st British Machine Vision Conference*, BMVC, 2020. Manchester, England. Acceptance Rate: 29%. DOI: BMVC-2020/0059

[3] **Gongbo Liang**, X. Wang, Y. Zhang, and N. Jacobs. “Weakly-Supervised Self-Training Breast Cancer Localization." In *Engineering in Medicine and Biology Society (EBMC), IEEE International Conference on.* IEEE, 2020. Montréal, Canda. DOI: 10.1109/EMBC44109.2020.9176617

[4] Y. Zhang, **Gongbo Liang**, T. Salem, and N. Jacobs. “Defense-PointNet: Protecting PointNet Against Adversarial Attacks." In *Big Data, IEEE International Conference on.* IEEE, 2019. Los Angeles, USA. DOI: 10.1109/BigData47090.2019.9006307

[5] **Gongbo Liang**, X. Wang, Y. Zhang, X. Xing, H. Blanton, T. Salem, and N. Jacobs. “Joint 2D-3D Breast Cancer Classification." In *Bioinformatics and Biomedicine, IEEE International Conference on.* IEEE, 2019. San Diego, USA. DOI: 10.1109/BIBM47256.2019.8983048

[6] Y. Zhang, X. Wang, H. Blanton, **Gongbo Liang**, X. Xing, and N. Jacobs. “2D Convolutional Neural Networks for 3D Digital Breast Tomosynthesis Classification." In *Bioinformatics and Biomedicine, IEEE International Conference on.* IEEE, 2019. San Diego, USA. DOI: 10.1109/BIBM47256.2019.8983097

[7] **Gongbo Liang**, S. Fouladvand, J. Zhang, M. Brooks, N. Jacobs, J. Chen. “GANai: Standardizing CT Images using Generative Adversarial Network with Alternative Improvement.” In *Healthcare Informatics, IEEE International Conference on.* IEEE, 2019. Shenzhen, China. Acceptance Rate: 28%. DOI: 10.1101/460188

[8] **Gongbo Liang**, Q. Li, and X. Kang. “Pedestrian detection via a leg-driven physiology framework." In *Image Processing, IEEE International Conference on.* IEEE, 2016. Phoenix, USA. DOI: 10.1109/ICIP.2016.7532895

[9] Q. Li, **Gongbo Liang**, and Y. Gong. “A geometric framework for stop sign detection." In *Signal and Information Processing, IEEE China Summit and International Conference on.* IEEE, 2015. DOI: 10.1109/ChinaSIP.2015.7230403

Refereed Workshops

[1] **Gongbo Liang**, S. Lin, Y. Zhang, Y. Su, and N. Jacobs . “Optical Wavelength Guided Feature Learning for Galaxy Group Richness Estimation." In *Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS) Workshop on Machine Learning and the Physical Sciences*, 2020. Vancouver, Canada.

[2] **Gongbo Liang**, Y. Zhang, and N. Jacobs. “Neural Network Calibration for Medical Imaging Classification Using DCA Regularization." In *International Conference on Machine Learning (ICLM) Workshop on Uncertainty & Robustness in Deep Learning*, 2020. Vienna, Austria.

[3] X. Xing, **Gongbo Liang**, H. Blanton, M. Rafique, C. Wang, A. Lin, and N. Jacobs. “Dynamic Image for 3D MRI Image Alzheimer’s Disease Classification." In *2020 the European Conference on Computer Vision (ECCV) Workshop on BioImage Computing*, 2020. Glasgow, United Kingdom.

Refereed Abstracts

[1] **Gongbo Liang**, N. Jacobs, and X. Wang. “Training Deep Learning Models as Radiologists: Breast Cancer Classification Using Combined Whole 2D Mammography and Full Volume Digital Breast Tomosynthesis." In 105^{th} *Scientific Assembly and Annual Meeting of the Radiological Society of North America (RSNA).* Chicago, IL, Dec 2019. Oral Presentation.

[2] Y. Zhang, **Gongbo Liang**, N. Jacobs, and X. Wang. “Unsupervised Domain Adapta-tion for Mammogram Image Classification: A Promising Tool for Model Generaliza-tion." In4thAnnual Scientific Conference on Machine Intelligence in Medical Imaging of the Society for Imaging Informatics in Medicine. Austin, TX,Sep 2019. Oral Presentation.

[3] **Gongbo Liang**, J. Zhang, M. Brooks, J. Howard, and J. Chen. “Enhancing Radiomic Features of CT Images using Generative Adversarial Network with Alternative Improvement.” In *AMIA Annual Symposium.* San Francisco, CA, Nov 2018. Poster presentation.

[4] **Gongbo Liang**, N. Jacobs, and X. Wang. “Breast Cancer Classification Using Combined Whole Mammography and Digital Breast Tomosynthesis.” In *2019 Markey Cancer Center Research Day.* Lexington, KY, May 2019. Poster presentation. (Award poster)

[5] **Gongbo Liang**, X. Wang, and N. Jacobs. “Evaluating the Publicly Available Mammography Datasets for Deep Learning Model Training.” In *2019 SBI/ACR Breast Imaging Symposium.* Hollywood, FL, Apr 2019. E-poster presentation.

[6] **Gongbo Liang**, J. Zhang, M. Brooks, J. Howard, and J. Chen. “Radiomic Features of Lung Cancer and Their Dependency On Ct Image Acquisition Parameters." In *59th Annual Meeting and Exhibition of American Association of Physicists in Medicine (AAPM).* Denver, CO, Jul 2017. Oral presentation.

[7] **Gongbo Liang**, J. Zhang, M. Brooks, J. Howard, and J. Chen. “Do Lung Tumor Image Features Depend on CT Acquisition Parameters.” In *American Associa-*

tion of Physicists in Medicine (AAPM) Ohio River Valley Spring Educational Symposium. Lexington, KY, April 2017. Oral presentation.

Honor and Awards

Apr 2020	Outstanding PhD Student, Computer Science Department, University of Kentucky
Mar 2020	One journal article was featured in “Inconsistent AI: Deep learning models for breast cancer fail to deliver after closer inspection" by Michael Walter, *AI in Healthcare*
Nov 2019	One research project was mentioned in “RSNA 2019 to offer a look at progress of AI and DBT" by Louise Gagnon, *AuntMinnie.com*
May 2019	Markey Cancer Center Research Day Second Place Awarded Poster Title: Breast Cancer Classification Using Combined Whole Mammography and Digital Breast Tomosynthesis
Apr 2016	WKU Student Research Conference Computer Science Session Winner Title: Pedestrian Detection Using Line Segments
May 2012	Robert J. Wurster Scholarship

Teaching

Spring 2021†	CSC316	3D Game Enging Design
Fall 2020†	CSC101	The World of Code
	CSC546/746	Artificial Intelligence
	CSC550/750	Graphics Programming
Fall 2019	CS275 Lab	Discrete Mathematics
Spring 2017	CS221 Lab	First Course in Computer Science for Engineers (Matlab)
Fall 2016	CS221 Lab	First Course in Computer Science for Engineers (Matlab)

Professional Activities

Reviewer

Since 2020	IEEE Transactions on Geoscience and Remote Sensing (TGRS)
Since 2019	IEEE Winter Conference on Applications of Computer Vision (WACV)
Since 2018	American Medical Informatics Association (AMIA) Annual Symposium
Since 2017	IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)
2020	International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)
2019	Journal of Applied Clinical Medical Physics (JACMP)
2018	Journal of Bioinformatics and Computational Biology
2016, 2018	Neurocomputing

Membership

Since 2016	Member of IEEE
Since 2016	Member of IEEE Young Professionals
Since 2020	Member of IEEE Engineering in Medicine and Biology Society

References

Available on request.