Computer-aided Detection of Breast Cancer in Digital Tomosynthesis Imaging Using Deep and Multiple Instance Learning

Mina Yousefi

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Computer Sciences) at

Concordia University

Montréal, Québec, Canada

May 2018

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mina Yousefi**

Entitled: **Computer-aided Detection of Breast Cancer in Digital Tomosynthesis Imaging Using Deep and Multiple Instance Learning**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Computer Sciences)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Amin Hammad

_____ External Examiner
Dr. Farida Cheriet

_____ Examiner
Dr. Tien D. Bui

_____ Examiner
Dr. Nawwaf Kharma

_____ Examiner
Dr. Sudhir Mudur

_____ Supervisor
Dr. Adam Krzyżak

_____ Co-supervisor
Dr. Ching Y. Suen

Approved by _____
Dr. Lata Narayanan, Chair
Department of Computer Science and Software Engineering

July 09, 2018 _____
Dr. Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Computer-aided Detection of Breast Cancer in Digital Tomosynthesis Imaging Using Deep and Multiple Instance Learning

Mina Yousefi, Ph.D.

Concordia University, 2018

Breast cancer is the most common cancer among women in the world. Nevertheless, early detection of breast cancer improves the chance of successful treatment. Digital breast tomosynthesis (DBT) as a new tomographic technique was developed to minimize the limitations of conventional digital mammography screening. A DBT is a quasi-three-dimensional image that is reconstructed from a small number of two-dimensional (2D) low-dose X-ray images. The 2D X-ray images are acquired over a limited angular around the breast.

Our research aims to introduce computer-aided detection (CAD) frameworks to detect early signs of breast cancer in DBTs. In this thesis, we propose three CAD frameworks for detection of breast cancer in DBTs. The first CAD framework is based on hand-crafted feature extraction. Concerning early signs of breast cancer: mass, micro-calcifications, and bilateral asymmetry between left and right breast, the system includes three separate channels to detect each sign. Next two CAD frameworks automatically learn complex patterns of 2D slices using the deep convolutional neural network and the deep cardinality-restricted Boltzmann machines. Finally, the CAD frameworks employ a multiple-instance learning approach with randomized trees algorithm to classify DBT images based on extracted information from 2D slices. The frameworks operate on 2D slices which are generated from DBT volumes. These frameworks are developed and evaluated using $5{,}040$ 2D image slices obtained from 87 DBT volumes. We demonstrate the validation and usefulness of the proposed CAD frameworks within empirical experiments for detecting breast cancer in DBTs.

"*It would not be much of a universe if it wasn't home to the people you love.*"

Prof Stephen Hawking

Acknowledgements

Joining the Centre for Pattern Recognition and Machine Intelligence (CENPARMI) at Concordia University was one of the best decisions of my life. I have to thank Dr. Suen and Dr. Krzyzak for their support and for many things that I learned from them. I started my research on deep learning models as the fantastic new learning model that is inspired by the human brain. Dr. Suen introduced me to the research group in the SAP company to work on the digital breast tomosynthesis image modality. This topic for me was very worthwhile, enabling me to research on the detection breast cancers automatically using machine learning algorithms as the most common cancer in women. Additionally, I have to thank Yanbin Lu, John Ellenberger, Richard H. Moore, and Daniel B. Kopans for helping me on this work during the first year of the project.

I would thank my family for their support in my studies. My parents and my parent-in-law helped me with all tough situations along the way. Finally, I thank my husband Mohammad Reza whose patience with my work and his loving support helped me to finish this research.

Dedicated to my beloved husband.

Contents

List of Figures

List of Tables

List of Abbreviations

2D	**T**wo **D**imensional
3D	**T**hree **D**imensional
CAD	**C**omputer **A**ided **D**etection
ROI	**R**egion **O**f **I**nterest
DCNN	**D**eep **C**onvolutional **N**eural **N**etwork
DBT	**D**igital **B**reast **T**omosynthesis
DM	**D**igital **M**ammography
GPU	**G**raphics **P**rocessing **U**nits
MIL	**M**ultiple **I**nstance **L**earning
MLO	**M**ediolateral **O**blique
CaRBM	**C**ardinality **R**estricted **B**oltzmann **M**achine
AT	**A**nscombe **T**ransformation
TPS	**T**hin **P**late **S**pline
RBST	**R**ubber **B**and **S**traightening **T**ransform
LOG	**L**aplacian **O**f **G**aussian
FFDM	**F**ull **F**ield **D**igital **M**ammography
NN	**N**eural **N**etwork
RBM	**R**estricted **B**oltzmann **M**achines
MRF	**M**arkov **R**andom **F**ield
DBN	**D**eep **B**elief **N**etwork
MI-RF	**M**ultiple **I**nstance **R**andom **F**orest
MI-SVM	**M**ultiple **I**nstance **S**upport **V**ector **M**achine
ROC	**R**eceiver **O**perating **C**haracteristic

AUC	**A**rea **U**nder the ROC **C**urve
FN	**F**alse **N**egative
TN	**T**rue **N**egative
FP	**F**alse **P**ositive
FP	**T**rue **P**ositive
s.t.	**S**ubject **T**o

Chapter 1

Introduction

Breast cancer is the most common cancer among women in many countries and the third most common cancer in Canada. According to the Canada breast cancer society, $26{,}300$ women would be diagnosed with breast cancer, which is 25% of all new estimated cancer cases in Canadian women in 2017 [1]. Four major cancer types in Canada are lung, breast, colorectal and prostate cancer. The lung cancer causes the most cancer deaths and breast cancer is the third most common cancer diagnosed. "Canadian females are more likely to develop breast cancer than any other cancer, with 1 in 8 females expected to develop breast cancer in their lifetime."(Quoted from [1]). All these figures show that more efforts are required to diagnose breast cancer at the early stage of cancer.

Breast cancer as a disease is the result of uncontrolled cellular propagation that invades healthy cells and forms tumors. The malignant tumors spread to other parts of the body and destroy nearby tissues. Breasts are composed of fatty, fibrous and glandular tissues (See Figure 1.1). Breast cancer is usually developed in the breast lobules (containing milk glands) or ducts, which produce and carry the milk to the nipple.

Over $32{,}000$ deaths from breast cancers have been avoided in Canada from 1988 to 2012. A reason for the reduction of the mortality rate from breast cancer is breast screenings. The most common breast imaging modalities are mammography, magnetic resonance imaging, ultrasound breast imaging, breast computed tomography, and digital breast tomosynthesis. Among these imaging modalities, two-dimensional (2D) X-ray screen-film, and full-field digital mammography modalities have been widely used for cancer screening. These imaging modalities have

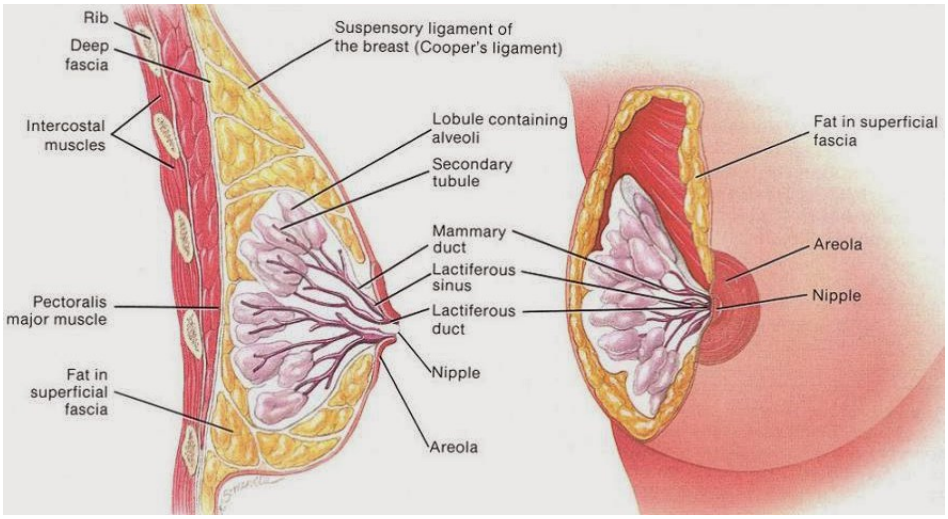


FIGURE 1.1: Breast anatomy (Extracted from [82]).

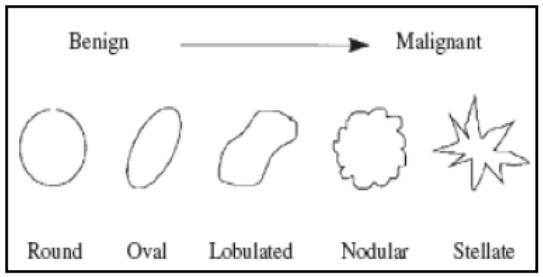


FIGURE 1.2: The level of malignancy increases from left to the right following the change of shape from circular to star (Extracted from [20]).

reduced breast cancer mortality rates over the course of 30 years [63]. Several significant cancer cues appear at the early stage of breast cancer in X-ray images. These early signs of breast cancers include spiculated mass, micro-calcifications, bilateral asymmetry, and architectural distortion. Among these four, in this thesis, we focus on the detection of masses, micro-calcifications, and bilateral asymmetry in digital breast tomosynthesis images.

A mass is a central space-occupying lesion which has variable density. Masses with regular shapes (circular and circumscribed oval) are benign cases. Malignant masses have shapes more like the star shape (see Figure 1.2). A spiculated mass as the primary indicator of breast cancer has been identified as a cluster of barbed tissue.

Calcium deposits such as calcite, aragonite, calcium, and apatite in the breast can be the early sign of breast cancer. According to the size of breast calcifications in an X-ray image, there are

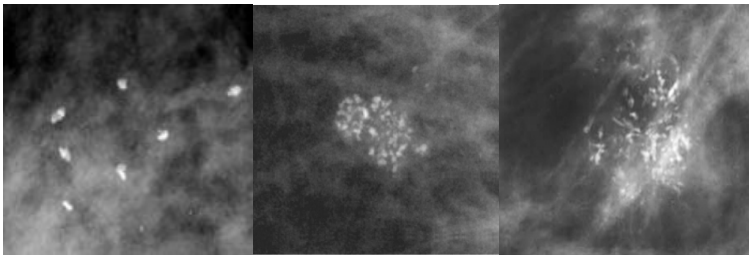


FIGURE 1.3: Examples of the variety of shapes of micro-calcification clusters mammograms (Extracted from the DDSM dataset [39]).

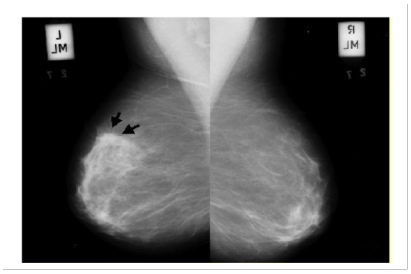


FIGURE 1.4: Example of the bilateral asymmetry between left and right breasts (Extracted from [84]).

two types of calcium deposits macro-calcification, and micro-calcification. A macro-calcification appear as large white dots (size $> 500\ \mu m$), or dashes in an X-ray image and they are almost harmless. Micro-calcifications are the tiny calcium dots (size $< 500\ \mu m$), and they may appear nearby cells that very quickly separated. A cluster of micro-calcifications according to the number of micro-calcifications and the shape of the cluster in radiography images may show the early sign of breast cancer (see Figure1.3).

Bilateral asymmetry is the asymmetry between right and left glandular (the functional tissue which includes lobules and ducts). The corresponding regions in a pair of normal tissues are similar but not for a mass [57] (see Figure 1.4). Radiologists measure the bilateral asymmetry by different factors to detect breast cancer in the early stage. [10]. There are two types of bilateral asymmetry: the global, and the focal asymmetry. When a higher volume of glandular tissues appeared in one breast compared to the same area in the other breast, it called as the global asymmetry. The focal asymmetry is determined as the circumscribed area of asymmetry in two views. The focal ones are the islands of healthy glandular tissues that are covered by the surrounding adipose tissues.

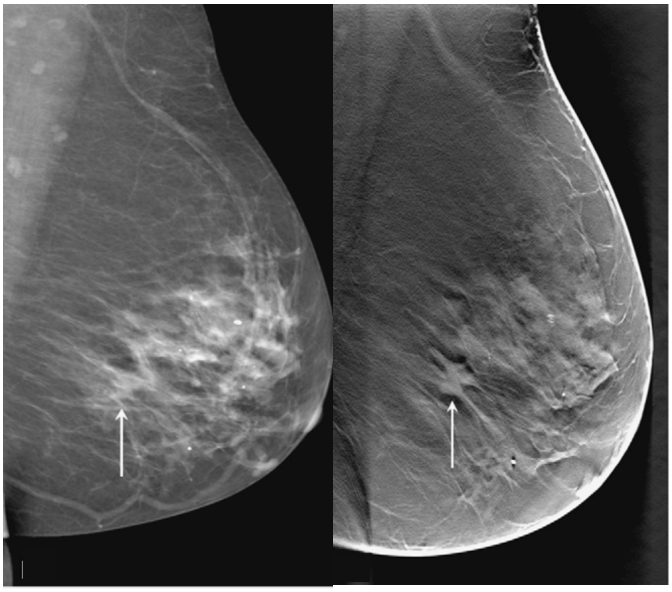


FIGURE 1.5: (a) A mammogram image shows the subtle architectural distortion (arrow). (b) A 2D slice of digital breast tomosynthesis image that shows a delicate irregular mass with the related architectural distortion (Extracted from [8]).

Architectural distortions are abnormal breast patterns. They occur due to the presence of abnormal tissues such as non-visible masses (see Figure 1.5). They can be observed in invasive breast carcinoma, or in radial scar condition.

The biopsy is the definitive method to assess the malignancy of breast tissue. In a breast biopsy, the sample of suspicious tissue is removed surgically, or with a needle core sampling procedure. Then the sample is processed to determine whether it is cancerous or benign. The breast imaging modalities allow the display of the breast abnormalities for the improvement in breast diagnosis and biopsy methods.

1.1 Breast Imaging Modalities

1.1.1 Magnetic Resonance Imaging (MRI)

The breast magnetic resonance imaging (MRI) device transmits radio frequency (RF) pulses into the patient and provides a stack of 2D slices images of the breast. A typical MRI screening of a

breast can show tissue with soft contrasts, cysts, enlarged ducts, hematoma, and ruptured breast implants [41]. An MRI image of a breast usually is incapable to detect breast abnormalities or tumors. So it is not a perfect tool for breast screening. This modality often is used with other tests for patients who are at high risk for breast cancer.

1.1.2 Breast Ultrasound

Breast ultrasound is a non-ionising imaging modality that uses sound waves to obtain the image from the breast. The sound waves with frequencies above the audible range reflect at breast tissues and create echoes. A computer connected to the ultrasound tool records the obtained echoes from breast tissues as an ultrasound image. The breast ultrasound images can show a solid malignant cancerous tumor, but no micro-calcifications. Like MRI, the breast ultrasound imaging is used as an additional breast screening tool.

1.1.3 Mammography

Mammography as a 2D X-ray radiographic imaging technique is the most common breast screening modality. A mammogram image is provided using a short-pulse X-rays whose frequency ranges among (1.7-3.6) * 10^{18} hertz. A digital detector as the image receptor recodes mammography images. An instance of a mammography system is given in Figure 1.6.

In a mammography test, the breast is pulled away from the body, compressed and placed between two glass plates of the mammography machine. It includes two views: the top-down view taken from above a horizontal compressed breast as cranial-caudal (CC), and the mediolateral oblique (MLO) view that captured from the side of a diagonal compressed breast. Digital mammography has the advantages of low cost and low dose x-ray radiation that make it the best suitable breast screening for the broad population.

Despite the mammography advantages, e.g., low-dose X-ray energy usage, the ability to detect early signs of breast cancer, the method suffers from some limitation. The high breast compression might hide many angular breast information, especially for dense breasts. A thick piece of breast tissue might conceal the cluster of micro-calcifications. For instance, assume one line of

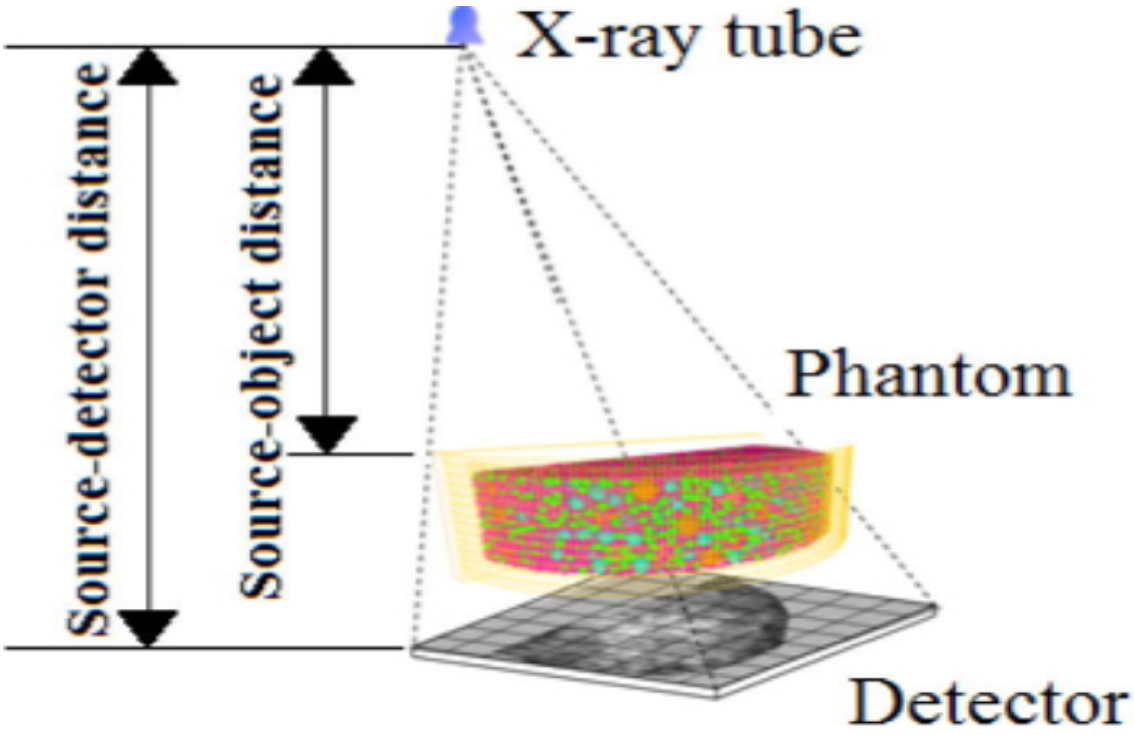


FIGURE 1.6: The simulated structure for a mammography image (Extracted from [11]).

calcification dots lies on the diameter of a cube in the breast, in digital mammography (DM) this diameter has appeared as a single dot. The radiologist may miss a micro-calcification cluster as an essential sign of breast cancer.

Providing a three-dimensional (3D) screening modality can resolve some current problems in mammography breast screening.

1.1.4 Breast Computed Tomography (Breast CT)

The breast CT as an X-ray imaging modality was invented to reduce mammography limitations by providing 3D X-ray images. This system takes hundreds of 2D X-ray images by spanning 360^o of the X-ray tube around a breast, then a 3D reconstructed image of the breast is generated from 2D X-ray images. Although the breast CT imaging removes the overlapping issue in the 2D X-ray images, it is an expensive test, and patients absorb high radiation dose (20-50 times more significant than a mammography test radiation dose). Moreover, the spatial resolution size

for a breast CT is much bigger than the one used for micro-calcifications detection. These are disadvantages of the breast CT modality.

1.1.5 Digital Breast Tomosynthesis (DBT)

The DBT as a new X-ray based on the limited angle imaging modality was proposed in the U.S. in 2003, and over the past decade, DBT has been developed in several researching labs. A high resolution, 3D and low radiation X-ray dose make DBT as an attractive alternative to breast CT [56].

Although there have been various DBT systems, they follow the same instruction for DBT screening. Most available DBT systems are comprised of an X-ray tube, a mechanical moving arm to shoot X-ray around the breast, compression paddle, an exposure control, a fixed image detector, and an object-support table [41], [56].

The compression paddle is placed on top of the breast to compress the scanned breast for minimizing the motion blur for the image. A DBT system provides a 3D reconstructed image from the scanned breast, so the tight breast compression is not necessary. The detector is fixed on the bottom of the breast. The tube path and the chest-side of the detector are set in the perpendicular orientation to the detector [56]. During the test, the X-ray tube moves around the breast over the limited angular ranges, acquires a series of projection views (PVs) like low-dose mammogram images. The range of angles and the number of projections usually varies between 20^o and 50^o, and 9 and 30, respectively. Figure 1.10 shows an example of a DBT system. A computer connected to the scanning machine provides a 3D image from the PVs using the application of 3D reconstruction techniques will be explained in chapter 2.

A DBT as quasi-3D mammography or an extension to the mammography modality allows doctors to look through layers of breast tissues. Compared to breast computed tomography, DBT can reduce mammography issues at an affordable cost. It also delivers much lower radiographic radiation dose to the patient, in which it uses only a fraction of the mammogram dose X-ray radiation to expose a projection-view of a breast, and the total dose remains equal to mammography X-ray dose or a little more than it [62], [34]. In each DBT test, a few X-ray projection

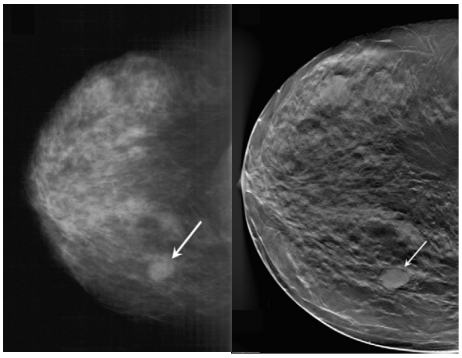
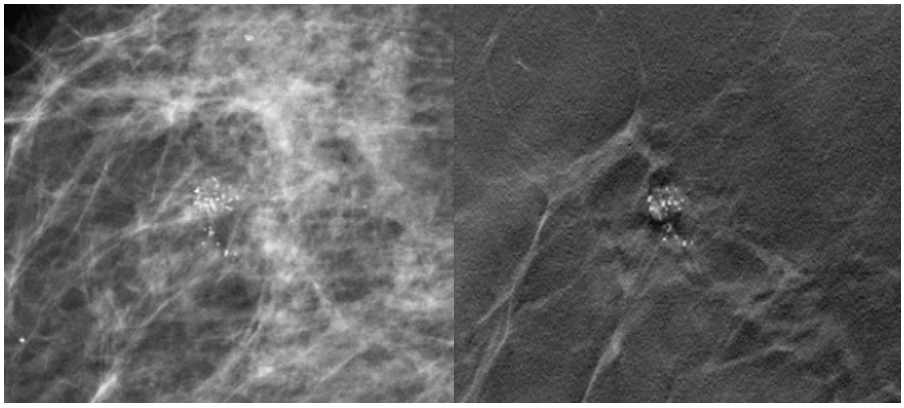


FIGURE 1.7: (a) A digital mammogram shows a mass that obscured, (b) a 2D slice from DBT that shows spiculated mass that is not obscured margin (Extracted from [8]).

views are provided that result in avoiding the risk of radiation-induced cancer for patients. Since a DBT provides a semi 3D imaging modality, it can eliminate the overlapping tissue problem of mammography tests, and facilitate visibility of malignant lesions. Moreover, the tight breast compression is not necessary during a DBT test. Note that these benefits have been achieved in DBT test by the radiation dose that remains in the range of mammography radiation dose. The DBT screening often is superior to digital mammography to detect breast cancer (See Figure 1.7, 1.8).

Despite DBT advantages, the challenges are encountered by DBT imaging. Insufficient 2D PV data causes the uncertainty in the reconstruction image. The processing time for providing DBT is high for instance, the time for the reconstructed 3D image. Finding the optimal trade-off between the X-ray dose that absorbed by the patient and image quality. In a DBT test, the radiologist has to review 50 to 80 2D X-ray images for a DBT test instead of the four ones in a mammography test. Thus developing new computer-aided detection (CAD) algorithms for assisting radiologists in evaluating DBT is an essential task.

Automatic breast cancer detection in medical imaging has been the critical studied topic for more than three decades. To develop the CAD framework for detecting breast tumors in medical images numerous subtasks are required for an accurate detection (knowledge of tumor patterns,



(a) **(b)**

FIGURE 1.8: (a) A cluster of calcifications in the mammogram, (b) the same area in a 2D slice of a DBT (Extracted from [8]).

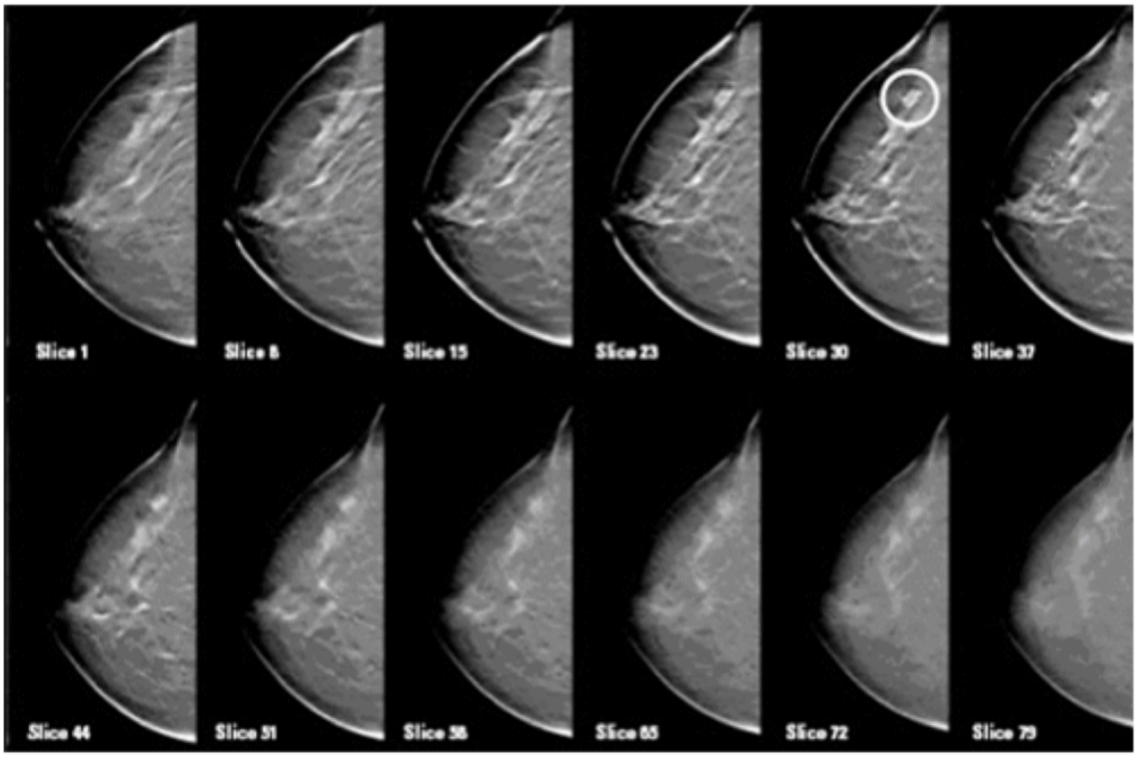


FIGURE 1.9: Some 2D slices of a DBT that show the breast cancer (the second left image in the top row), while it is hidden in other slices (Extracted from [79]).

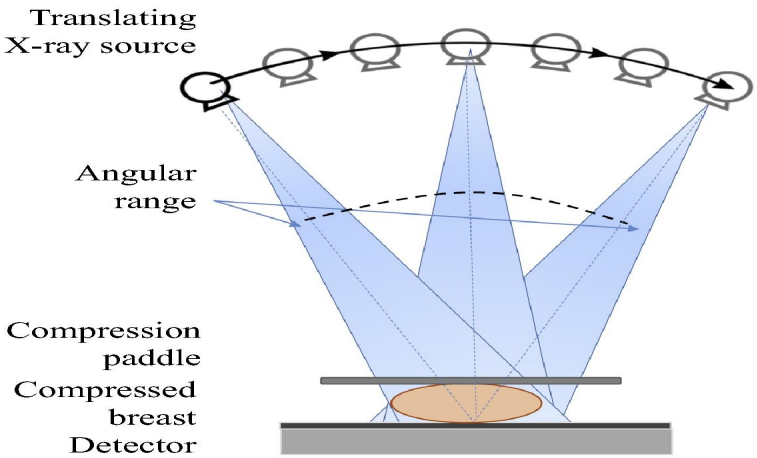


FIGURE 1.10: The X-ray system to capture DBT images [72].

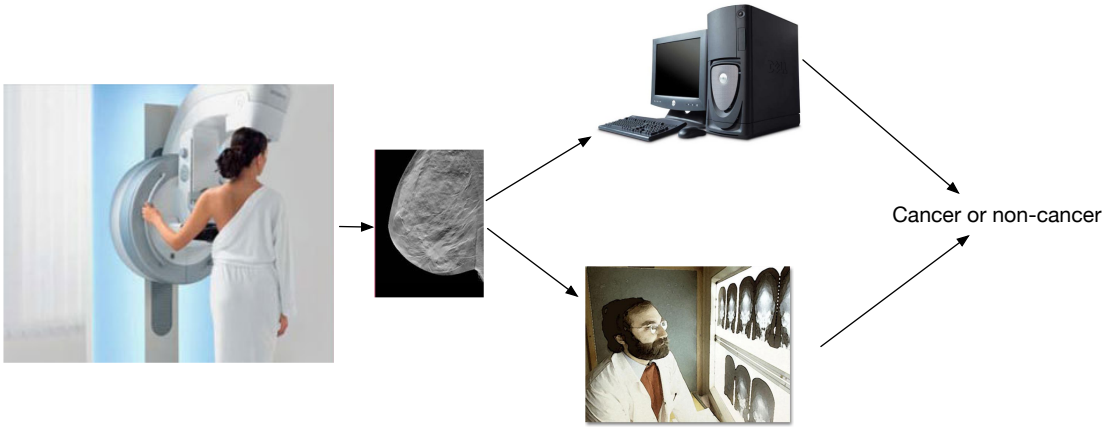


FIGURE 1.11: A simple graphical overview of a CAD framework of breast cancer in DBT images.

image preprocessing, the region of interest extraction, classification). Many studies have shown that the concurrent CAD framework in a DBT test can reduce the reading time by keeping the accuracy [9]. Moreover, unlike computer systems, a human manual approach often results in more unexpected errors; hence, the development of precise CAD systems plays an essential role in the further diagnosis of DBT. Typically a CAD session comprises three steps (see Figure 1.11). First, the radiologist looks into screens, then the CAD framework analyzes the image, and the radiologist looks into the results provided by the CAD and makes the final decision.

1.2 Contributions

In this dissertation, our work focuses on exploring novel CAD frameworks for detecting breast cancer in DBT. We develop three different CAD systems for detecting masses in DBT. The first method utilizes the hand-crafted features to capture information from images. The second and third frameworks use deep deep cardinality-restricted Boltzmann machines and deep convolutional neural networks to extract features from images automatically. All three CADs employe the same multiple-instance learning method to classify DBT samples based on the obtained information from 2D slices. Moreover, we introduce two algorithms for detecting 1) clustered micro-calcification, and 2) bilateral asymmetry in DBT.

- The core ideas in this study include:
  - usage of the weakly supervised learning classification algorithm called multiple-instance learning in CAD systems for mass detection in DBT;
  - development of two deep learning models based on Cardinality restricted Boltzmann machines and the convolutional neural network for representing features of DBT slices;
  - a weighting scheme based on the bank of Gabor filters for Initial mass detection in 2D slices;
  - a reregistration algorithm in bilateral asymmetry detection algorithm on DBT.
- Implementation toolboxes for CAD frameworks based on the multi-layer tomosynthesis images. Each CAD system has own and the separate toolkit that is the implemented in Python and C++.

  (1) The toolkit for speculated mass detection algorithms includes:
  - implementation of practical and straightforward preprocessing methods for removing quantum noise, artifacts, and pectoral muscles from DBT;
  - the accurate method for extracting mass candidate regions on DBT slices, and two segmentation algorithms;

- feature extraction module includes some of the features are designed for this study, and others are adopted from the published studies;
- data augmentation algorithms based on rigid and non-rigid techniques for deep learning models;
- the deep cardinality restricted Boltzmann machines used in the second CAD system for mass detection in DBTs;
- the deep convolutional neural networks used for detecting masses in the third CAD system;
- multiple-instance learning algorithms include five classifications:
  - the multiple-instance learning with randomized trees that combines the advantage of randomized trees classifiers and flexibility of multiple-instance learning;
  - four support vector machines-based multiple instance learning algorithms.

(2) The toolkit for bilateral asymmetric detection in DBT images includes:

- the accurate thin-plate spline registration algorithm;
- comparison algorithms to measure the similarity of aligned images from left and right breast is estimated based on the intensity and feature values.

(3) The micro-calcifications detection toolkit includes:

- the Laplacian of Gaussian filter as the detecting kernel;
- the post constraint methods to apply constraints on the calcification candidates.

The work in this thesis has been published in the following papers:

- M. Yousefi, A. Krzyżak, and C. Y. Suen. Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning. Computers in Biology and Medicine, volume 96, pages 283 – 293, 2018. [104].

- Y. Lu, M. Yousefi, J. Ellenberger, R. H. Moore, D. B. Kopans, A. Krzyżak, and C. Y. Suen. 3D tomosynthesis to detect breast cancer. In Handbook of Pattern Recognition and Computer Vision, chapter 2.10, pages 371–393. World Scientific, 5th edition, 2015 [58].
- M. Yousefi, Y. Lu, C. Y.Suen, and A. Krzyżak. Computer aided diagnosis images to detect breast cancer. Computer Methods in Biomechanics and Biomedical Engineering (CMBBE), Symposium, Montreal, Canada, 2015 [103].
- M. Yousefi, A. Krzyżak, and C. Y. Suen. Convex cardinality restricted boltzmann machine and its application to pattern recognition. In Advances in Artificial Intelligence, volume 8436, pages 369–374, 2014 [102].

1.3 Structure of the Thesis

Chapter 2 presents an overview of different computer-aided detection (CAD) frameworks for breast cancer detection in digital breast tomosynthesis images. This chapter presents a literature review of 3D reconstruction algorithms on DBTs and a review of the available CAD frameworks for detecting breast cancer in DBT. Chapter 3-4 are the main chapters which represent the architectures of CAD frameworks for breast cancer detection in DBTs. Chapter 3 presents the hand-crafted, feature-based CAD system, and two algorithms for bilateral asymmetry and micro-calcification detection. We show the deep cardinality-restricted Boltzmann machines- and deep convolutional neural networks-based frameworks in chapter 4. This chapter will is followed by the representation of the multiple-instance-, random forest, support vector machines, based classifiers. In chapter 5, we explain our dataset that we use throughout the thesis. The rest of this chapter represents the results of experiments for three CAD frameworks with five different multiple-instance learning are evaluated using different measurement techniques. In the final chapter, the thesis ends with the discussion, conclusion and an outline of future work.

Chapter 2

Literature Survey

The role of DBT as the primary screening modality has not be determined. The number of issues must be addressed before popular using tomosynthesis for the primary clinical test. For instance, the number PVs per image, angular spacing, the radial arc of the X-ray, and its dose remain topics of research of the DBT device. The list of DBT units approved by the Food and Drug Administration (FDA) for marketing use is provided in Table 2.1.

Although DBT is a 3D volume imaging technique and gives the dimensional information, the complete 3D details on the breast cannot be reconstructed. So one of the major subjects for DBT is the improvement of DBT image quality. One of the approaches to reduce current problems in DBT breast screening is the development of automatic detection techniques for assisting the radiologists in diagnosis. Several research groups have developed CAD systems to detect breast cancer and other breast diseases in DBTs. Based on the nature of DBT images, a CAD framework can be developed either based on 2D data [70], [58], or on volume data [80], [92].

In this chapter, we first discuss the different significant drawbacks of DBT images that are intrinsic to the DBT design and how to eliminate them. Next, we provide a literature review of existing CAD models for detecting breast cancers in DBT images.

TABLE 2.1: DBT systems approved by FDA for marketing (Extracted from the FDA homepage)

DBT systems	Year
GE Senographe Pristina with Digital Breast Tomosynthesis	2017
Fujifilm ASPIRE Cristalle with Digital Breast Tomosynthesis (DBT)	2017
Siemens MAMMOMAT Inspiration with Digital Breast Tomosynthesis (DBT)	2015
GE SenoClaire Digital Breast Tomosynthesis (DBT) System	2014
Hologic Selenia Dimensions Digital Breast Tomosynthesis (DBT) System	2011

2.1 Reconstruction Approaches in DBT

Reconstruction algorithms in DBT can impact the whole clinical DBT performance [80]. DBT is a limited angle, cone beam breast CT imaging modality. Compared to breast CT, DBTs suffer from incomplete projection views and data truncation, because the X-ray tube and the detector do not rotate completely around the breast. The fundamental issue for reconstructing a DBT is the null space problem. This problem is inherited from the nature of DBT image that limited PVs are taken. In the DBT reconstruction process, only slices that are perpendicular to the direction of the projection are obtained. So some regions in the reconstructed image are unknown [41]. To address this issue, utilizing the anatomical priors in the reconstruction procedure may be helpful.

The reconstruction algorithms in breast CT imaging modality that are dependent on the completeness of data are not suitable for DBTs. So new reconstruction algorithms have been investigated for DBT images. Though many researchers have achieved the progress in this domain, more efforts are required to solve this reconstruction problem. In this section, we review the literature on DBT reconstruction approaches.

2.1.1 The Shift-and-add Reconstruction

The first DBTs have been provided by a clinical system called full-field digital mammograms (FFDM) system. The FFDM acquired nine projections over an X-ray source angular of range 40^o. The simple shift-and-add (SAA) algorithm reconstructed 3D DBT from projection views images. The SAA algorithm shifts pixels that are parallel to the X-ray tube motion and adds them in the reconstructed planes. This algorithm determines the geometry from projections using radio-opaque fiducial points. With the improvement in DBT systems, DBTs acquired from more projections, and these systems utilized more accurate reconstruction algorithms.

2.1.2 Reconstruction by Fourier Methods

The Fourier approach is the two-step reconstruction method, application of the Fourier transform on the parallel projections from an object and reconstruction from the transform. One way of providing the reconstructed image is the application of the direct Fourier method and the filtered back-projection (FBP) on transformed images. This method suffers from the error that comes from the nature of Fourier transform as the image points at higher frequencies become sparser.

The blurring-out filtered back-projection (FBP) reconstruction works like the Fourier transform reconstruction, but it uses the ramp-filtered in the frequency domain. The FBP algorithm filters out low frequencies and passes the high frequencies. So it oversampled the Fourier domain at lower radial frequencies, resulting the blurring is minimized [60] [81] [56]. Micro-calcifications visualize more clear in DBT images that reconstructed using the FBP algorithm when compared with reconstructed images using the SAA algorithm.

2.1.3 Statistical Iterative Reconstruction Approaches

The maximum likelihood reconstruction algorithm offers a reconstruction model using photon-counting signals [52]. The likelihood function is defined as the probability of the set of observing projection views and the estimator of the reconstruction object. The posterior distribution is determined using the Bays rule that combines the likelihood function and image prior. The

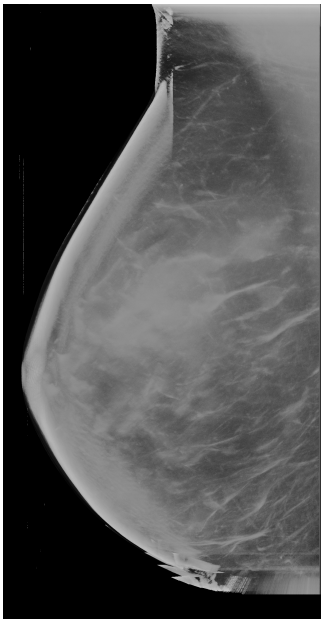


FIGURE 2.1: A 3D reconstructed DBT image.

objective of the reconstruction algorithm is to find the optimal reconstruction estimator from the obtained projections, in which the incident of the X-ray and the current image parameters to maximize the posterior distribution. The likelihood function usually has been modeled using Markov random field models. The maximum likelihood algorithm provides for mass and micro-calcification visualizations better balance when compared with other reconstruction algorithms.

2.1.4 Algebraic Reconstruction Methods

The simultaneous algebraic and matrix-inversion tomosynthesis reconstruction algorithms are used to reconstruct 3D DBT [92]. These models provide a series of linear equations from PVs as the weighted sum of the incident and transmitted intensity. To reconstruct the image, these series of equations simultaneously are solved to obtain the corresponding incident values for PVs. The algebraic methods need fewer projections compared to FBP algorithm. These algorithms keep the sharpness of micro-calcifications and improve the contrast-to-noise ratio. They can provide 3D DBTs faster than maximum likelihood algorithm with comparable image quality. These models suffer from the significant computation time.

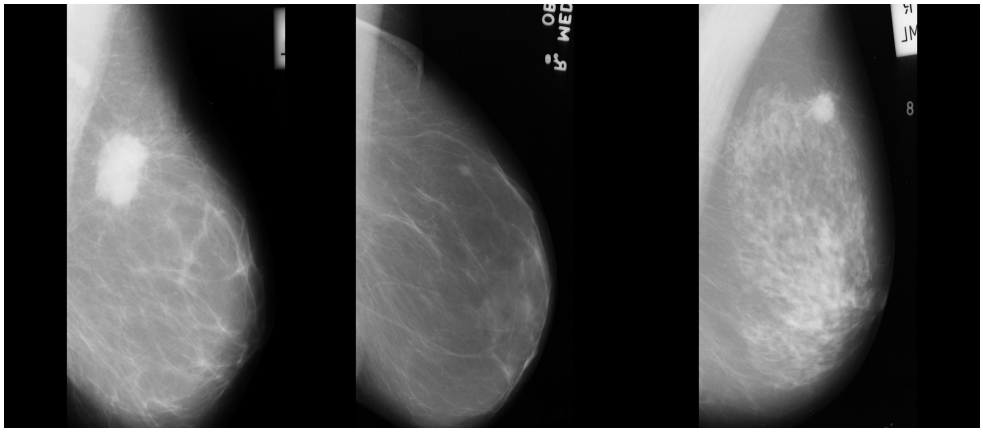
2.2 Analysis of DBT X-ray Images

2.2.1 Breast Density in X-ray Images

Breast density is one of the most critical image-based measurements for risk of the breast cancer. Usually, the higher density means, the higher breast cancer risk [4]. Radiologists approximate the breast density from their visual assessment of the mammogram images. The X-ray has a small range of wavelength (0.1pm to 10nm) with the corresponding energy (100 eV to 10 MeV). The X-ray with shorter wavelength and higher energy can interpolate the most materials. The amount of X-ray energy that used in a breast screening test varies for breasts with different density. The breasts consist of two component tissues: fibroglandular tissues, and fat. Fat has the lower X-ray attenuation than fibroglandular tissues, so the regions of fat appear darker in X-ray images [100]. For measuring the density of a breast, the difference between the appearance of the fibroglandular, and adipose tissue compositions are calculated based on qualitative, and quantitative factors. In an X-ray image, the brighter regions show the region of interests. For instance, dense, fibrous, and glandular appear lighter than adipose tissue [41]. The fibroglandular tissues appear as a shade of the gray to white, and fatty tissue appears darker [41]. One of the standard classification scheme is used for breast density, represents the density of a breast by the percentage value. It measures the breast appearance dense with the number of glandular tissues. A huge dense breast has the percentage of density value bigger than 75%. Using DBT images, we can make much more sufficient accurate estimation of breast density than mammogram images. On each slice separately the fatty tissues are segmented from the dense tissues, then the participation of each slice image would be combined. Figure 2.2 shows some mammogram samples which diagnosed spiculated mass with different density breasts from the mini-MIAS database [84].

2.2.2 DBT Image Quality

The quality of medical X-ray imaging can affect the performance of detection abnormalities tasks. Two factors are considered to measure the quality of a medical image. How the image is distinctive, and details of the object are visible in the image [41]. The quality of X-ray images



(a) **(b)** **(c)**

FIGURE 2.2: Three samples diagnosed spiculated mass: (a) Fatty, (b) Fatty-glandular; (c) Dense-glandular.

as an essential problem must be assessed and improved, before the application of any analytical algorithms.

Noise in X-ray Images

Noise as a random signal can hide the signal with the useful information. So typically in the first stage of image analysis procedure, the source of noise, and noise removal ways are investigated. The basics principles of all X-ray images from breast are the same, so they have similar image quality. Digital mammography includes the structure noise that it reduces in DBT and breast CT images. However, these 3D image modalities include out-of-plane artifact noise.

Quantum Noise

In the process of generating X-ray imagining, photons randomly distributed within the image. It means that the number of photons varies in various locations of the images. These randomly distributed photons in X-ray images cause the various contrast of X-ray images in the different area of the image. Increment X-ray radiation dose reduces the quantum noise, but the amount of the absorbed X-ray radiation by patients in tests is an essential factor in medical imaging. Each projection view for a DBT is generated using less than $\frac{1}{6}$ of the dose of two DM exposures. So it

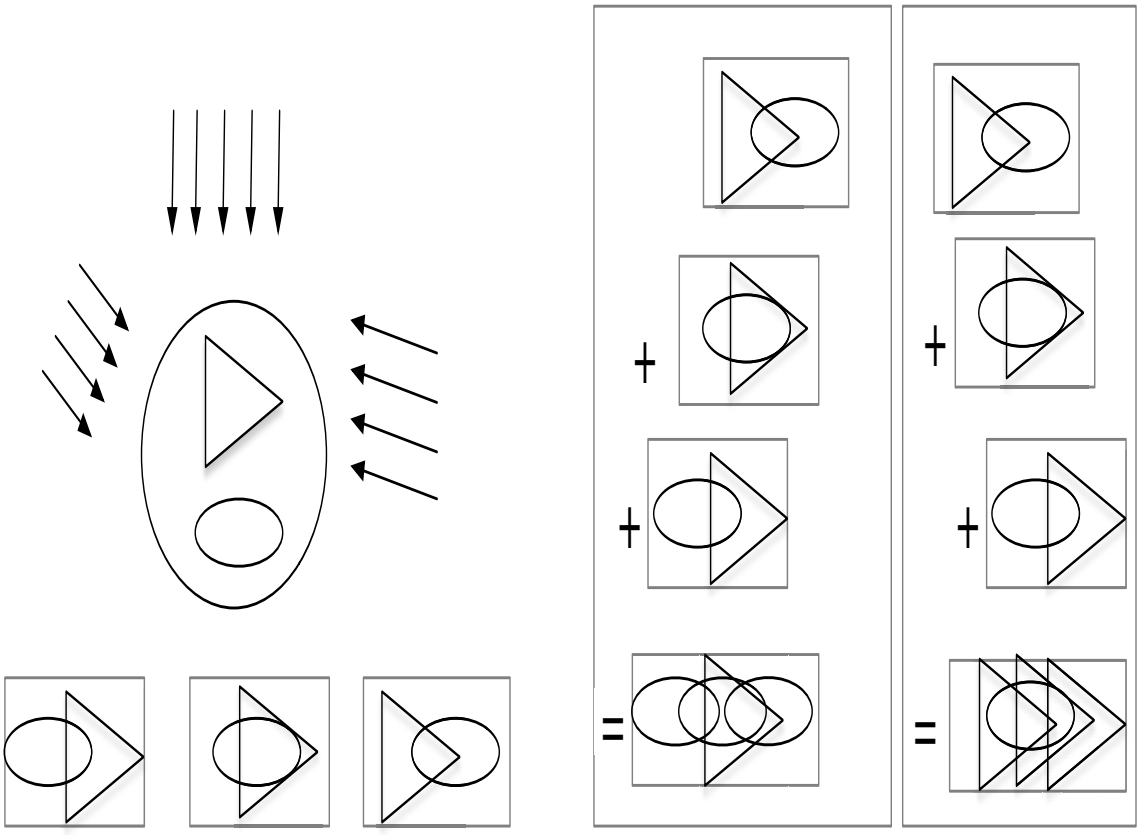


FIGURE 2.3: The three-projected views (left), two slices after reconstruction (right). (Inspired by the model in [41])

is expected the quantum noise in DBTs is more critical than in DMs. Hence, we should optimize the trade-off between removal of quantum noises of DBTs and X-ray radiation dose.

Out-of-plane Artifacts

This noise has appeared as some low-contrast shadows in the plains. The DBT reconstruction algorithm causes this noise. We can explain this noise with a simple reconstruction example using the shift-and-add reconstruction algorithm. Assume we take three X-ray images of the object which includes two patterns. As shown in Figure 2.3, the objects in reconstructed outputs using the SAA algorithm are blurred, while these patterns are more visible in the plane of interest. Stronger reconstruction algorithms can reduce out-of-plane artifacts in the reconstructed images [15]. Figure 2.4 shows out-of-plane artifacts noise in DBT samples.

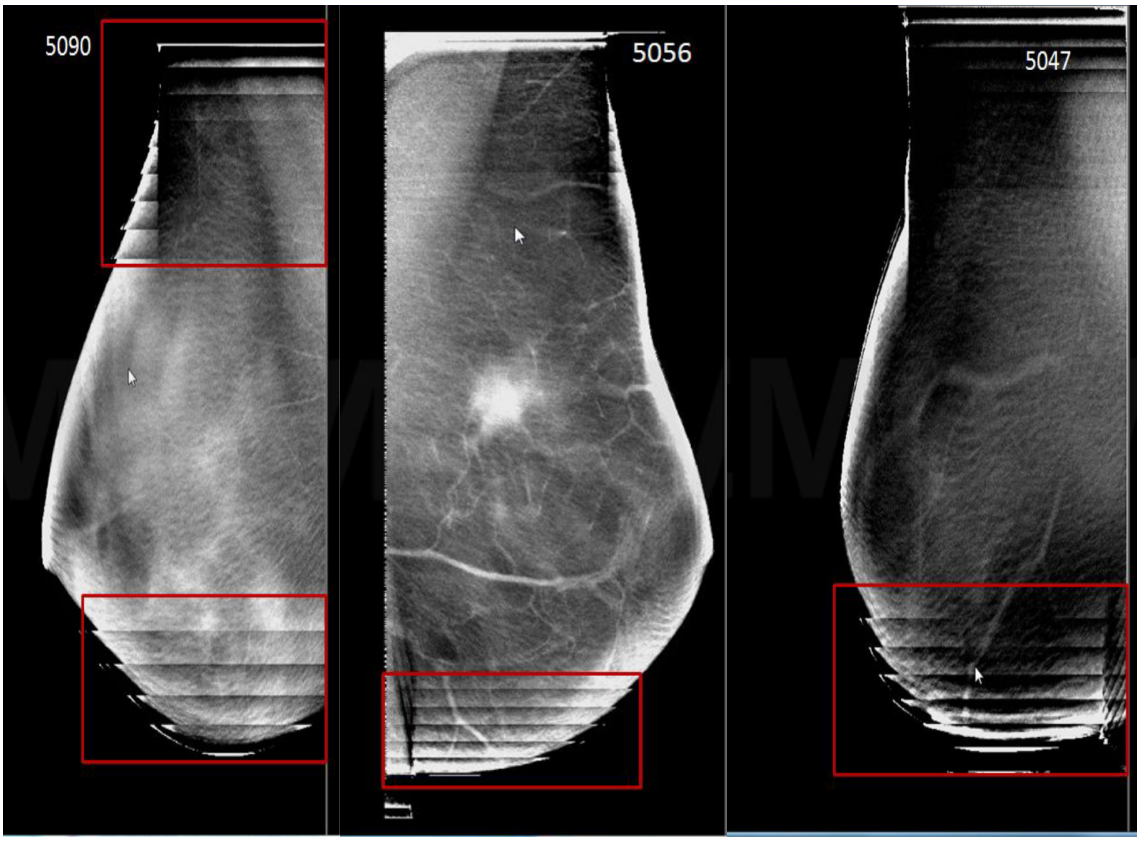


FIGURE 2.4: There are horizontal lines on the three sample DBT images that show the artifact noise on the DBT images.

2.3 Overview of Computer-aided Detection Frameworks on DBTs

Over the past two decades, many research groups have developed CAD systems for detecting breast abnormalities in X-ray images. A CAD system as the automated tool can localize and annotate lesions in images. The procedure of detecting breast tumors in medical images using automatic CAD systems typically comprises three steps (see Figure 2.5): lesion detection and segmentation, feature-extraction of the lesion, and lesion classification [69], [30]. In the feature-extraction stage usually, a group of hand-crafted features (e.g., statistical, morphology, fractal, etc.) are extracted from mass candidate regions. Additionally, in the last few years, deep learning-based frameworks have achieved outstanding performance in a variety of mammogram,

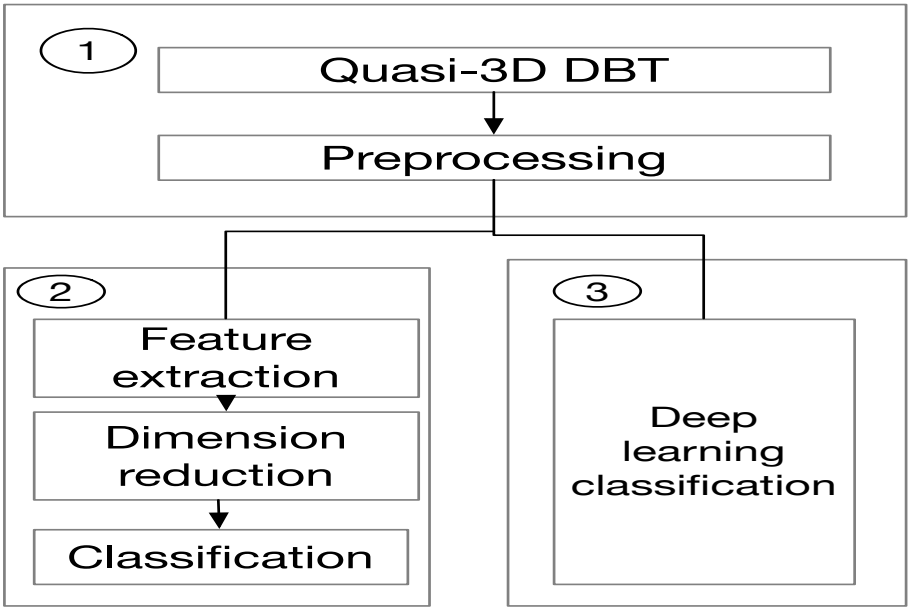


FIGURE 2.5: Flowchart of CAD systems using the hand-crafted features and deep learning approaches. Module (1): preprocessing, module (2): a CAD method using feature extraction and false-positive reduction and module (3): image classification and false-positive reduction using deep learning machine.

and DBT analysis approaches [22], [44], [85], [12], [76], [35]. These approaches extract natural features from data through self-learning in deep architectures.

In this section, we present the literature review of some existing CAD frameworks for detecting spiculated masses and micro-calcifications in DBTs.

2.3.1 Mass Detection Approaches in DBT Volumes

2.3.2 Hand-crafted Feature-based CAD Frameworks

Chan et al. [25] presented two CAD frameworks for mass detection on DBT volumes and PV images. The information obtained from DBT volumes and PV images were merged to find 3D mass likelihood scores in DBT images. The regions of interests were extracted from each slice of the DBT image using the breast boundary detection algorithm. Then the volume of interests (VOI) by a 3D region growing method was detected in which it uses the obtained areas from 2D slices. In the next stage, three types of features morphological, gray level, and texture feature

were extracted from VOIs. The linear discriminant analysis (LDA) classifier with the leave-one-case-out re-sampling training technique was used to identify masses in DBT volumes. The CAD framework on PV images, used the gradient filed algorithm to detect masses independently on PVs, in which mass candidates were identified as locations with high gradient convergence. The K means clustering approach was used to extract the object from the background. Three types of features, morphological, Hessian, and texture, were obtained from ROIs. The LDA classifier with stepwise feature selection technique was employed to detect masses. The mass likelihood images derived from the PVs are back-projected to the breast volume to calculate mass likelihood scores for 3D images. The averaging method merged two extracted mass likelihood from two CAD frameworks to find the 3D mass likelihood scores in DBT images.

Palma et al. [65] proposed the CAD scheme that comprises two channels for masses and architectural distortions detections in DBT volumes. In the mass detection channel, the fuzzy connected filters were applied to the mass candidate detection stage. The adaptive thresholding method extracted seeds from the fuzzy detection regions and conveyed them to the two-stage segmentation approach. In the feature extraction stage, the authors extracted entropy, contrast, homogeneity, etc. features. They used the support vector machine classification as the classifier. In the second channel, convergences were detected using a contrario approach. The false positive reduction step was followed by the 3D aggregation step to take the 3D information. Then statistical features were extracted, finally, the disjunctive fusion method merged two channels.

Wei et al. [95] represented a CAD framework that utilized the adaptive segmentation approach for mass detection in DBT. Mass candidates were selected from DBT images by the 3D gradient field approach. The 3D clustering approach was applied to mass candidates then it followed by the 3D active contour segmentation to provide VOIs. Three features: morphological, gray level, and texture features were extracted from VOIs. The LDA classifier was used to find the mass likelihood scores.

2.3.3 Deep Learning Based CAD Frameworks

Deep learning approaches especially, deep convolutional neural networks (DCNNs) have been successful in classifying medical images Carneiro et al. [22], Fotin et al. [35]. In deep learning approaches, the image patterns are recognized from the training data and incorporated into the millions of parameters Samala et al. [76].

Samala et al. [76] designed the DCNN for mass detection in DBT images. They used the transfer learning approach for pretraining the DCNN with four convolutional, and three fully connected layers [76]. In the learning procedure, they used both mammogram and DBT images. The augmented mammography ROIs data were utilized for pretraining the CNN model. After training the model on mammogram data, the weights of the first three convolutional layers were frozen; then the last convolutional layer and fully connected layers were trained using the DBT training ROIs. The total number of ROIs after data augmentation for mammography training data was $45{,}072$, and for DBT training set was $37{,}450$.

Fotin et al. [35] used the DCNN inspired by the AlexNet model [51] as a multi-layer architecture that consists of alternative convolutions and non-linearities which followed by the fully connected layers and the softmax classifier. They applied the deep model on 3D extracted regions of interests from DBT images to detect masses. They used the augmentation technique using random shifting, rotation, and scaling to enlarge the data. The DCNN was trained on patches sampled from mammography images including $1{,}864$ suspicious tissues and 339 lesions from slice-by-slice images of DBT volumes.

Kim et al. [47] proposed the DCNN based CAD system for detecting latent bilateral feature representation of masses in DBT volumes. The DCNN includes four convolutional and four subsampling layers and two fully connected layers. First, low-level features were extracted from the ROIs and the corresponding ROIs through the convolutional layers separately. Then low-level features have been combined in FC layers to represent the high-level bilateral features of the masses.

As the significant limitation of these systems, I like to mention that these systems utilize supervised learning classifications. Consequently, their CAD models need the labels for whole data in

datasets. Practically, this is a tough job for physicians to mark all ROIs of all slices from DBT images. So we need to investigate the proper weakly supervised learning classifications that are suited for DBT images.

2.3.4 Micro-calcifications Detection in DBT

There are several micro-calcification detection approaches can be categorized into CAD models based on DBT volumes and 2D PV images.

Micro-calcification Detection Approaches on DBT Slices

Bedea et al. [7] developed a noise removal method using the noise mask subtraction approach on 2D slices. In the micro-calcification process, a generated noise-mask was applied to the plane which more likely includes micro-calcifications. The selected plane was projected for all tomographic angles that used onto the plane of the DBT image. These projected views reconstructed a blurred image. Note that this blurred image was originated from the selected plane including noise that forms the noise-mask. After subtraction the noise-mask from the plane of interest, micro-calcification detection procedure took place on the de-noised slices.

Bernard et al. [14] introduced a method to measure the contrast of micro-calcifications using Mexican hat wavelet responses on each DBT slice. The process was based on the assumption that the spherical micro-calcification with radius R was placed in a breast of uniform attenuation coefficient. The difference of the mean gray level between the inside the focus slice (crossing the center of the sphere) and the neighborhoods of the slab was obtained by convolving the Mexican hat wavelet with the focus slice. The background noise estimation, 3D connected component algorithm to extract micro-calcifications, and 3D clustering of micro-calcifications also were used in this framework.

Sahiner et al. [73] introduced a CAD system to detect micro-calcifications in the DBT volumes. In the prescreening stage, DBT images were enhanced based on the contrast and their structures. Using the global and the region growing methods micro-calcification candidates were identified on images. The potential micro-calcifications initialized the dynamic clustering algorithm

within a 3D neighborhood. Then the rule-based classifier was used to reduce the number of false positives (FPs).

Samala et al. [75] presented the CAD framework for detection of micro-calcifications in DBT volumes. This system utilized multi-scale bilateral filter regularization to enhance DBT images during the reconstruction procedure. It improved micro-calcifications using the combined multi-scale Hessian response with enhancement-modulated calcifications response. The iterative thresholding method by the region growing was adapted to locate micro-calcification candidates. The dynamic clustering formed the clusters, and some criterions based on the size, the number of micro-calcifications and the contrast-to-noise ratio were applied to reduce the number of FPs.

Micro-calcification Detection Methods on Projection Views

Peters et al. [67] presented the projection views based approach for micro-calcifications detection in DBT. This method utilized the responses of PVs to the multi-scale wavelet filter to obtain micro-calcifications candidates. The membership value for each candidate was calculated with creating a fuzzy contour using the multi-level threshold process. A fuzzy particle map was used for the segmentation, and the aggregation operator combined information over the complete set of PVs. Then the results were mapped to 3D locations.

Reiser et al. [71] proposed a general micro-calcification detection algorithm. In this scheme, each PV image was treated as a mammogram image. The morphologic de-nosing method was followed by the difference-of-Gaussian (DoG) for detecting micro-calcification candidates. The local thresholding approach was used for the segmentation. All segmented images were back-projected into the 3D feature volume. The maximum-intensity projection approach identified calcification candidates. Then 3D breast volumes were reprojected to PV images. Linear discriminant analysis (LDA) and the round-robin method were used for micro-calcifications detection.

Wei et al. [96] developed a 2D approach for micro-calcifications detection using PV images. Each PV image was enhanced using the contrast-to-noise ratio (CNR) filter. The iterative thresholding method detected micro-calcifications candidates. Micro-calcifications were decomposed

with Hermite-Gaussian (HG) and Laguerre-Gaussian (LG) basis functions. The channelized Hotelling model was trained to obtain the multichannel response for each micro-calcifications on PV images. 2D responses were back-projected into the 3D volume by the coincidence counting method. Finally, a dynamic clustering method identified micro-calcifications clusters in DBT images.

Samala et al. [77] designed a DCNN to differentiate micro-calcifications from the planar PV image. Because the image patterns were simple, the structure of the DCNN model was simple with only two convolution layers, and a fully connected layer.

HO introduced the epipolar curves approach to detect micro-calcifications [41]. The cluster of micro-calcifications in DBT that appeared as a single dot in the 2D coordinate frame was called the epipolar curve. The clustering algorithm using the Hough transform was designed to find all corresponding points which were generated from one micro-calcification. The detection algorithm used Markov random fields and belief propagation to gather all the detection results in PVs. The cluster of micro-calcifications was reconstructed from the detected micro-calcifications in PVs.

Chapter 3

Hand-crafted, Feature-based CAD Framework to Detect Breast Cancers in DBTs

In this chapter, we introduce the hand-crafted, feature-based CAD framework that we developed for DBT images. In this CAD model, we focus on the detection of three early signs of the breast cancer: the presence of spiculated masses, clustered micro-calcification, and bilateral asymmetric through the three-channels detection scheme. Figure 3.1 shows the overview of the framework.

3.1 Methodology

Before the application of any analytical algorithms on images, a few problems related to image contrast and image noise must be addressed in the preprocessing module. This module of the CAD framework comprised of the noise removal, skin-line, and pectoral muscle removal stages. We also provided the option of contrast enhancement using the histogram equalization algorithm. Detection spiculated masses channel includes initial detection mass-like object using the bank of Gabor filters, extraction regions of interests (ROIs), hand-crafted feature extraction, and classification DBT images. In the micro-calcifications detection channel, we use Laplacian of Gaussian filter as the primary technique in this model, besides the foreground extraction, and

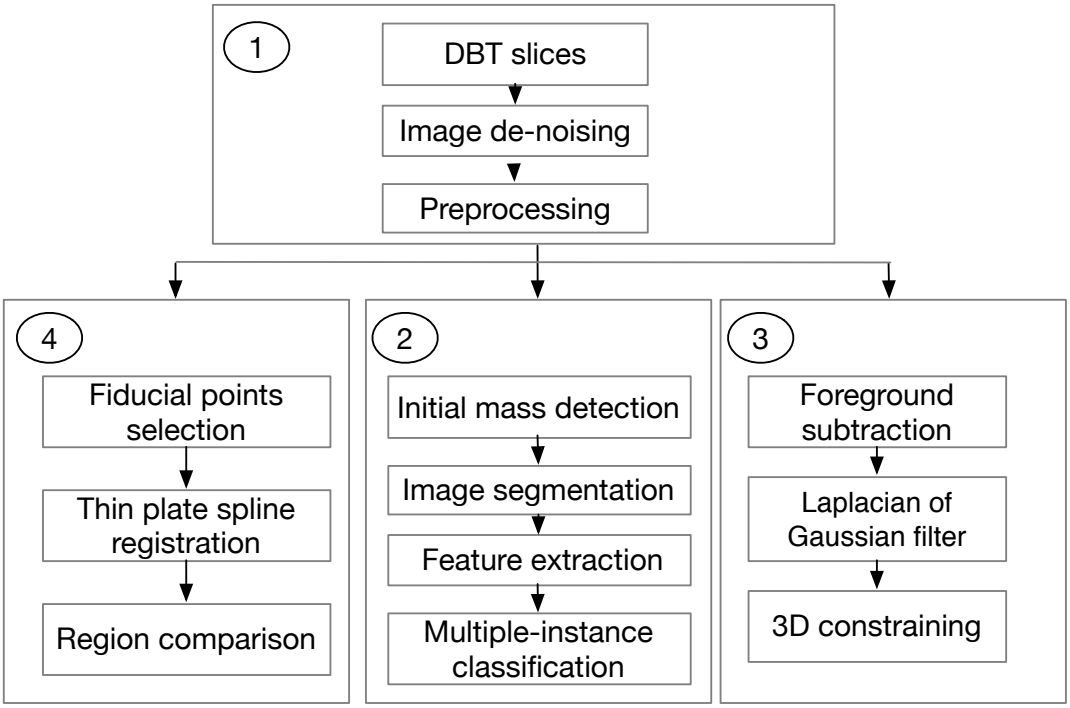


FIGURE 3.1: (1) The image preprocessing module, (2) mass detection channel, (3) micro-calcifications channel, and (4) bilateral asymmetry analysis channel.

post-processing steps are applied to trim the detected micro-calcifications. The third channel of this CAD framework is designed for identifying and analyzing bilateral asymmetry in DBT images. To investigate the asymmetry of two breasts, we compare each 2D slice image from the left breast by the same slice number from the right breast. This channel comprises the contour detection, the fiducial points selection, the image registration, and region comparison stages.

3.2 Mass Detection Channel

3.2.1 Preprocessing

3.2.2 Image De-nosing

As mentioned before, DBT images are provided by low dose X-ray radiation, and X-ray images are formed using the photon counting that follows the Poisson distribution given by:

$$P(Z = z|y) = \frac{y^z e^{-y}}{z!} \tag{3.1}$$

where Z is a Poisson random variable with mean y and z that represents photon counts. The signal-dependent quantum noise covers the DBT images. This noise that corrupts DBT image has Poisson distribution with dependent mean and variance. Theoretically, any algorithms can stabilize the variance of distribution will apply to eliminate noise. In this study, we employ the combination of Anscombe variance-stabilizing transformation (AT) and adaptive Wiener filter de-nosing method [93]. Anscombe showed that $\sqrt{z+\frac{3}{8}}$ is the nearest constant variance transformation for z with Poisson distribution [6]. The AT converts a random variable z with Poisson distribution into the variable with Gaussian distribution with zero mean and unit variance. The z variable is transformed using the AT given by:

$$f(z) = 2 * \sqrt{z + \frac{3}{8}} \tag{3.2}$$

The AT transforms the random variable into the one with a Gaussian distribution. Then, the adaptive Wiener filter is used to remove the noise in Gaussian distribution [45] from the transformed data. Afterward, the inverse transform of AT is applied to the de-noised data D to return the variance-stabilized and de-noised data in the original range. There are the variety of inverse transformation approximations [59] that are listed below. In this study, we use the exact unbiased inverse way to return the data to the original format.

- Numerical inverse:

$$I_A(D) = f^{-1}(D) = (\frac{D}{2})^2 - \frac{3}{8} \tag{3.3}$$

- The adjusted inverse:

$$I_B(D) = (\frac{D}{2})^2 - \frac{1}{8} \tag{3.4}$$

- The exact unbiased inverse:

$$E\{f(z)|y\} = 2\sum_{z=0}^{\infty}(\sqrt{z+\frac{3}{8}}.\frac{y^z e^{-y}}{z!}) \tag{3.5}$$

  where $E\{z|y\} = y \geq 0$ represents the true value of the signal.

- Maximum likelihood (ML) inverse:

$$I_{ML}(D) = \text{argmax}_y p(D|y) \tag{3.6}$$

$$p(D|y) = \frac{1}{\sqrt{2\pi\varepsilon^2}} \mathrm{e}^{-\frac{(D-E\{f(z)|y\})^2}{2\varepsilon^2}} \tag{3.7}$$

  where $\varepsilon^2 = E\left\{(D - E\{f(z)|y\})^2\right\}$

- minimum mean square error (MMSE) inverse:

$$I_{MMSE}(D, \varepsilon) = \text{argmax}_{\hat{y}} E((y - \hat{y})^2|D) \tag{3.8}$$

$$\hat{y} = \frac{\int_0^\infty p(D|y) y dy}{\int_0^\infty p(D|y) dy} \tag{3.9}$$

3.2.3 Pectoral Muscle Removal

Since our data are prepared in the MLO view, pectoral muscles (PMs) appear on images with the high intensity. The PMs are looking like a breast fibroglandular parenchyma (a breast tissue) in X-ray images. As we know, the intensity of mass candidate regions in X-ray images is higher than in normal breast tissues, and most mass candidate identification approaches utilize this property in their procedures. So identification and removal of PMs are critical prior steps to analyze DBTs. In this study, we use the Canny edge detector and the probabilistic Hough to detect PMs boundary, and the thresholding kernel to filter out the muscle area in 2D slices (see Figur 3.3).

3.2.4 Skin-line Removal

The skin-lines (boundary of the breast) always do not have any beneficial information for analyzing X-ray images. The intensity of these breast boundaries appear high, and they can affect subsequent image processing. We use a simple thresholding approach to get rid of them from images. In our dataset, image intensity values are distributed between $4{,}000$ and $6{,}500$, while the background value of the image is zero. We use two thresholds the lower threshold $2{,}500$,

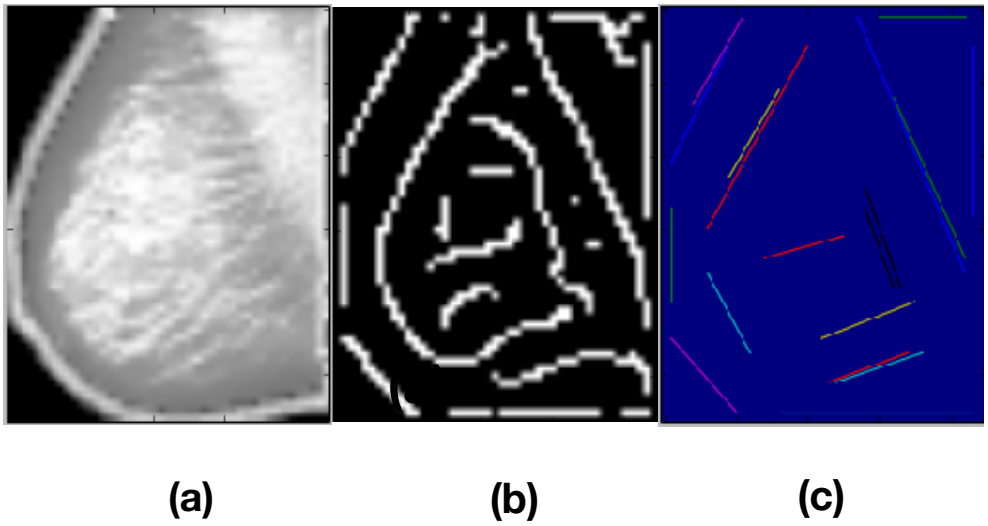


FIGURE 3.2: (a) The original image. (b) After application of the Canny edge detection algorithm, and (c) the probabilistic Hough transform of the input image in (a).

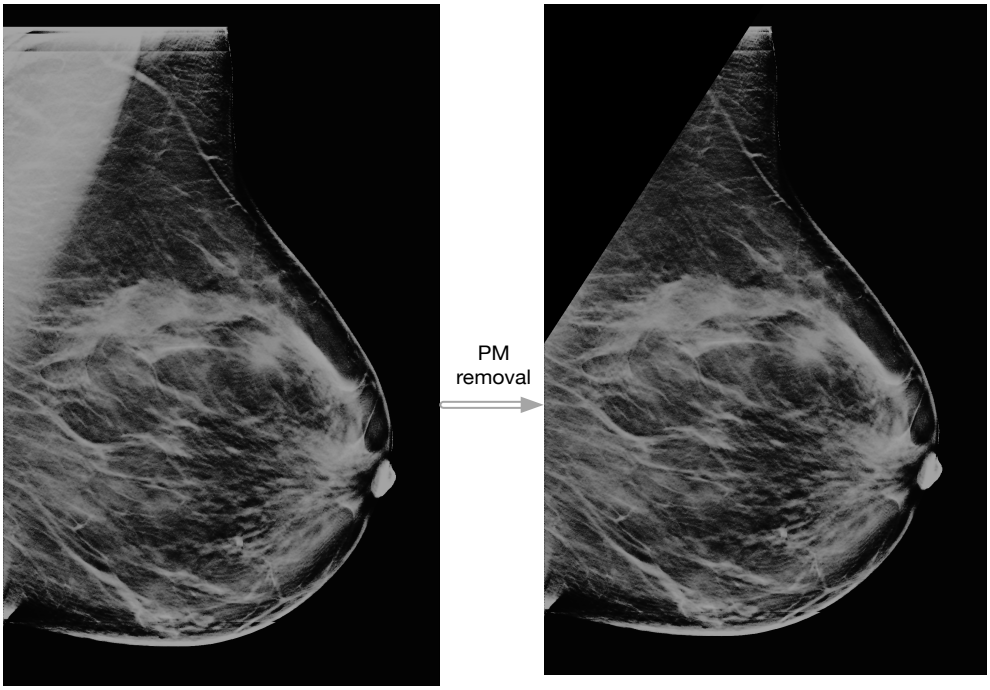


FIGURE 3.3: The 2D slice image before and after the PM removal stage.

and the higher one 7,500 to eliminate the breast boundary and artifacts. The two binary masks are dilated to remove the boundary and artifacts. We use the dilating element mask as disks with

the of 15, and $30mm$, respectively (see Figure 3.4). After the skin-line removal stage, Otsus segmentation is applied on images to resize 2D slices to 256×256.

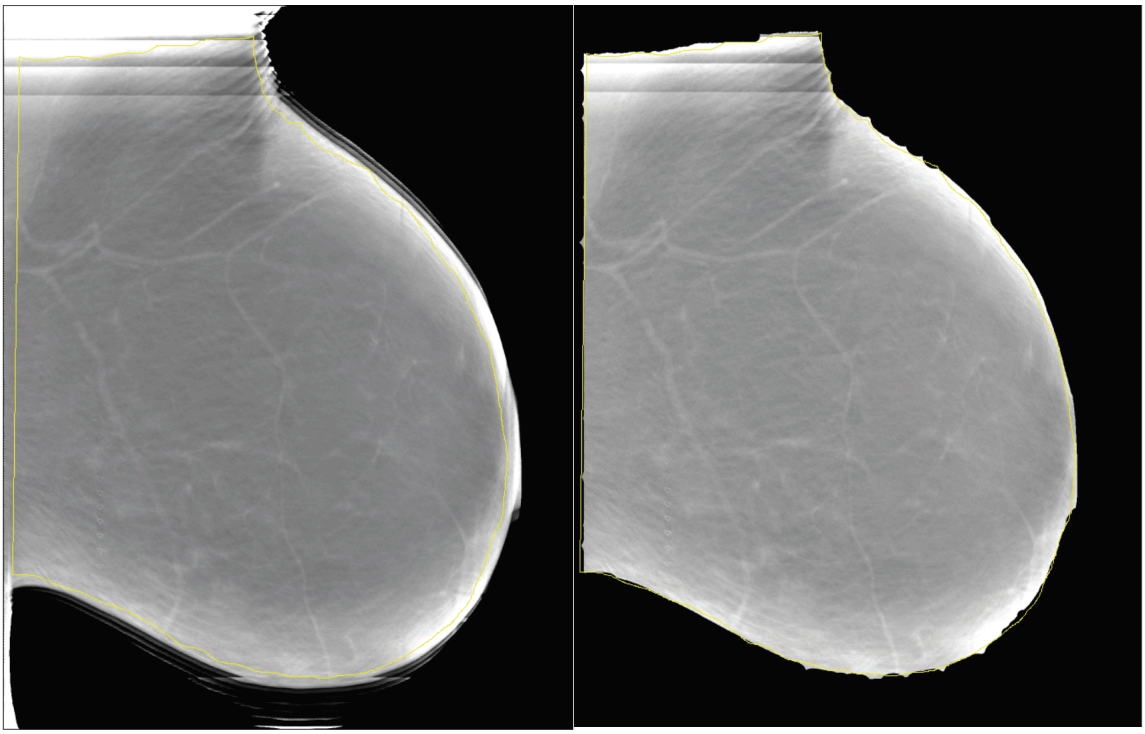


FIGURE 3.4: 2D sample image before, and after boundary and artifact removal.

3.2.5 Mass Regions Detection Approach

Initial Mass Detection

The regions of interest (ROI) are required to be specified for extracting features that identify malignancy breast lesions. Given the morphological characteristic of spiculated masses, which is a lump of tissue with spikes or points on the surface, we tend to detect objects with higher than average intensity and spike-like shapes. We use the bank of Gabor filters with multi-frequency, multi-orientation. Gabor filters are the promising filters for texture analysis. 2D Gabor filters are defined as the product of a complex sinusoidal wave and Gaussian function 3.10. The idea behind of using Gabor filters on our dataset is that a centered blob with the spike-like surface responds strongly to a Gabor kernel with the specific frequency (frequency corresponds to the spike frequency), and its rotating orientations kernels (see Figure 3.5). Thus, in a particular area

of the image, if the response to a kernel and its rotating siblings are higher than the neighborhood areas, it is highly possible that a spike-like blob object exists in this area. Specifically, a few steps are involved in localizing the coordinate of suspicious spiculated mass: the first step is Gabor transformation. In the second step, the voting procedure analyzes the filters responses. The voting procedure works based on a weighted sum of response magnitude and response orientation. The magnitude vote of a pixel for a Gabor kernel is the summation of the pixel responses to all orientations of the kernel. The orientation voting is calculated based on the ratio of a kernel orientation response as follows: if the rate of the kernel response to the sum of all kernel orientation responses of the image is higher than the threshold, the orientation parameter in the voting score is increased by one. If the response to a Gabor kernel with the specific frequency (frequency corresponds to the spike frequency), and its rotated siblings are highest in its neighboring areas, and higher than the threshold, the area is kept as the mass candidate region (see Figure 3.6).

$$\mathrm{G}(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2})\exp(\mathrm{i}(2\pi\frac{x'}{\lambda} + \psi)), \tag{3.10}$$

$$x' = x\cos\theta + y\sin\theta \quad \text{and} \quad y' = -x\sin\theta + y\cos\theta$$

where λ, θ, ψ, σ and γ are the wavelength of the sinusoidal factor, the orientation of the normal to the parallel strips, the phase offset, the standard deviation of the Gaussian envelope and the individual ratio, respectively [101].

Mass Segmentation

To fully capture the properties of each ROI, we apply the interactive segmentation method to ROIs. This segmentation algorithm utilizes a principled deep random field model and the graph cut algorithm [49]. Using the standard pairwise Markov random field (MRF), we define the segmentation approach for mass candidate regions. For an ROI, the MRF graph is the pairwise (V, E) where V and E are sets of nodes (pixels of the image) and edges (adjacent pixels pairs), respectively. The nodes of the graph are the binary variables in which the value of a binary variable $x_i \in X$ is $(1/0)$ depending on whether an image pixel belongs to the foreground or the

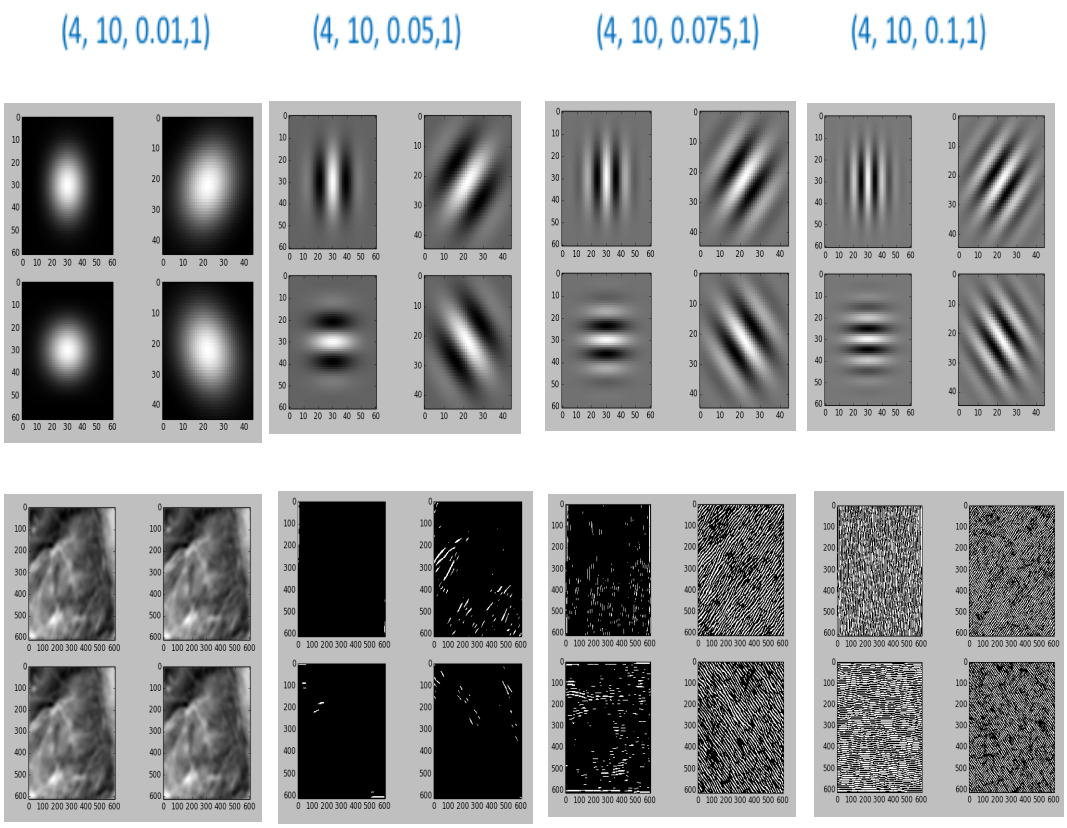


FIGURE 3.5: The images in the top row are Gabor kernels with parameters (number of orientations=4, scale, frequency, gamma). The corresponding responses to the Gabor kernels are in the bottom row.

background, respectively. The pairwise potential $\psi(x_i, x_j)$ is defined on the neighboring binary variable nodes of the graph. Then we set the objective of the segmentation approach so that it finds the optimal X which maximizes the posterior distribution of the MRF field:

$$\begin{aligned} \mathrm{P}(X|I) &= \frac{1}{Z} e^{-\mathrm{E}(X)} \\ \text{s.t.} \quad \mathrm{E}(X) &= \sum_{i \in V} \phi(x_i) + \sum_{(i,j) \in E} \psi(x_i, x_j), \end{aligned} \tag{3.11}$$

where I, Z, $\phi(x_i)$ are the image, the normalizing constant, and a unary potential function, respectively. The unary potential function shows the likelihood score of the pixel i being in the image foreground. For the optimization procedure, the energy function $\mathrm{E}(X)$ in e.q 3.11 is minimized over X. Pairwise potential function in 3.12 penalizes different labels of neighboring pixels

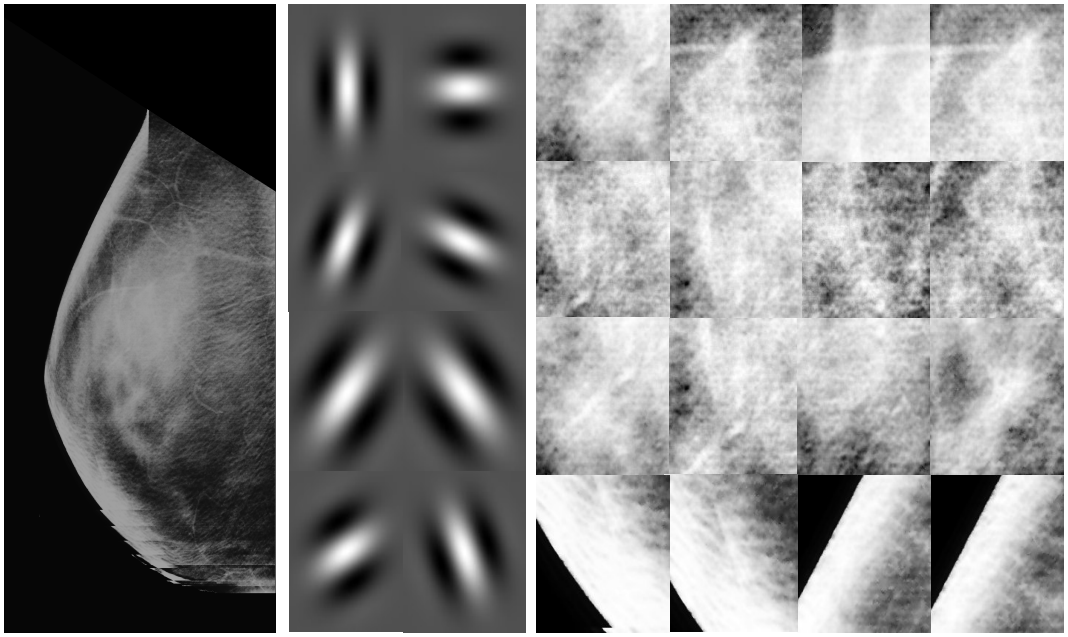


FIGURE 3.6: (b) Gabor kernel with 8 orientations applied on a sample image. (c) Mass candidate patches from the image in (a) which produce responses to Gabor filters in (b) higher than the threshold 16.4.

(4-neighbor or 8-neighbor pixels). This penalty encourages smooth segmentation and simultaneously discourages long object boundaries. Consequently, the short-boundary bias problem appears in the segmented image. Since there are massive objects boundaries pixels which do not share the same label as their neighbors, the short boundary bias is a major issue for medical image segmentation. The cooperative graph cut approach addresses this problem, in which this method replaces the domain of the pairwise potential in 3.12 by the diversity transitions between boundary pixels, and their neighbor's sets. In spite of massive objects boundaries pixels that do not share the same label as their neighbors, the number of label transitions is constant, and the diversity of changes between boundary pixels and their neighbors is limited.

$$\begin{aligned} E(X) &= \sum_{i \in V} \phi(x_i) + \sum_{g \in G} \Psi_g(x) \\ \text{s.t.} \quad \Psi_g(x) &= \mathrm{F}\Big(\sum_{(i,j) \in G} \psi(x_i, x_j) \Big) \end{aligned} \tag{3.12}$$

F function, and G are the nondecreasing concave function as the lower envelope of multiple linear functions with matching templates in [49], and the diversity boundary object pixels set, respectively.

As explained above, the boundary energy of the cooperative cut algorithm is a sum of thresholded submodular functions, while the boundary energy in the graph cut is a sum of so fully modular functions. Relaxing thresholds in a cooperative cut algorithm in the boundary energy make the submodular functions. If we use many edges, a submodular over the edges of the corresponding region gives us a discount on the cost of the long cut [43]. We provide a comparison example that shows the outputs of the cooperate cut and graph cut algorithms on the ROI (see Figure 3.8).

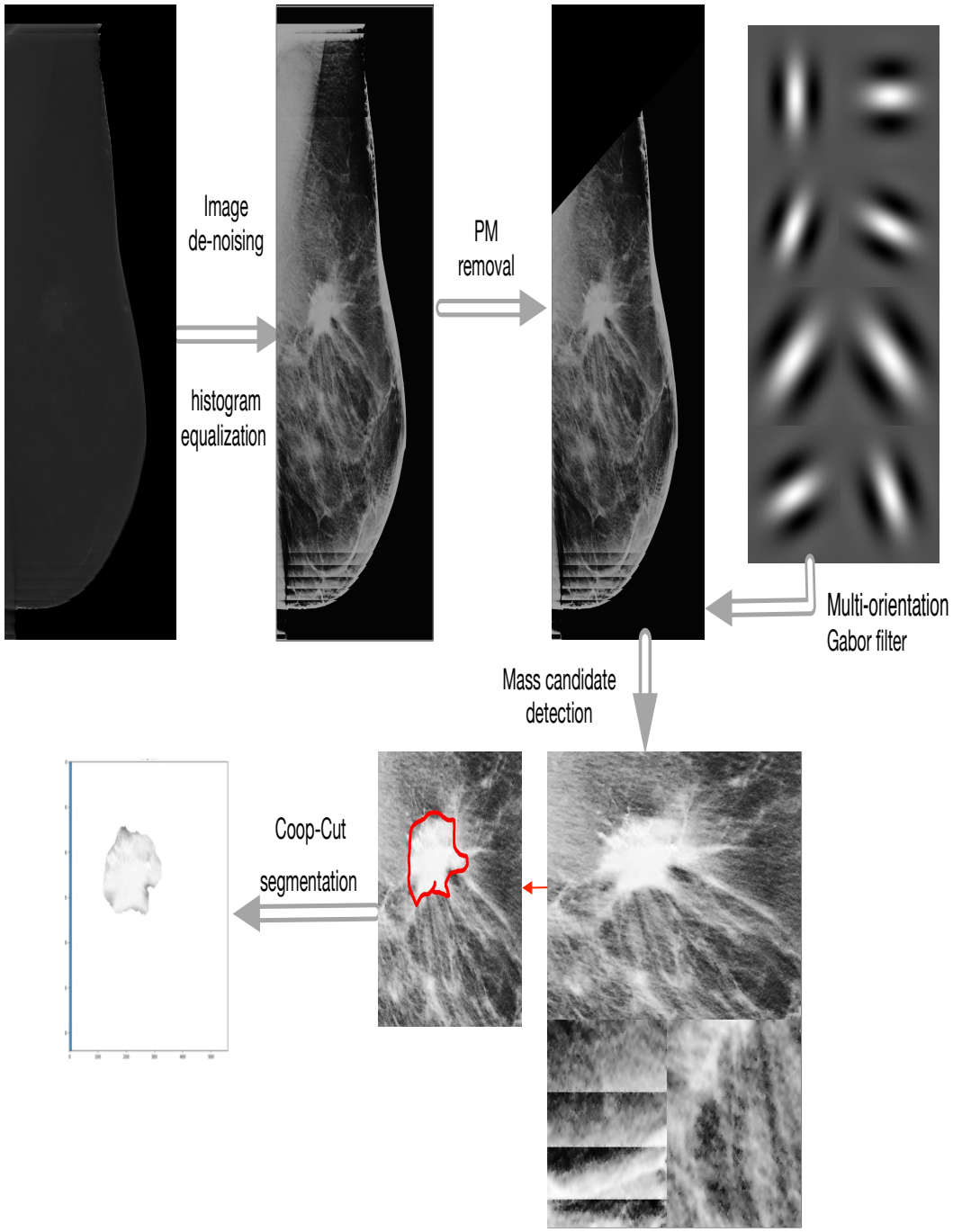


FIGURE 3.7: Image de-noising, PM removal, the region of mass candidate detection, and segmentation stages on a sample image.

3.2.6 Feature Extraction

How to efficiently design the reliable features is the challenge for all the breast cancer researchers. The extracted features should include critical information from images so that breast

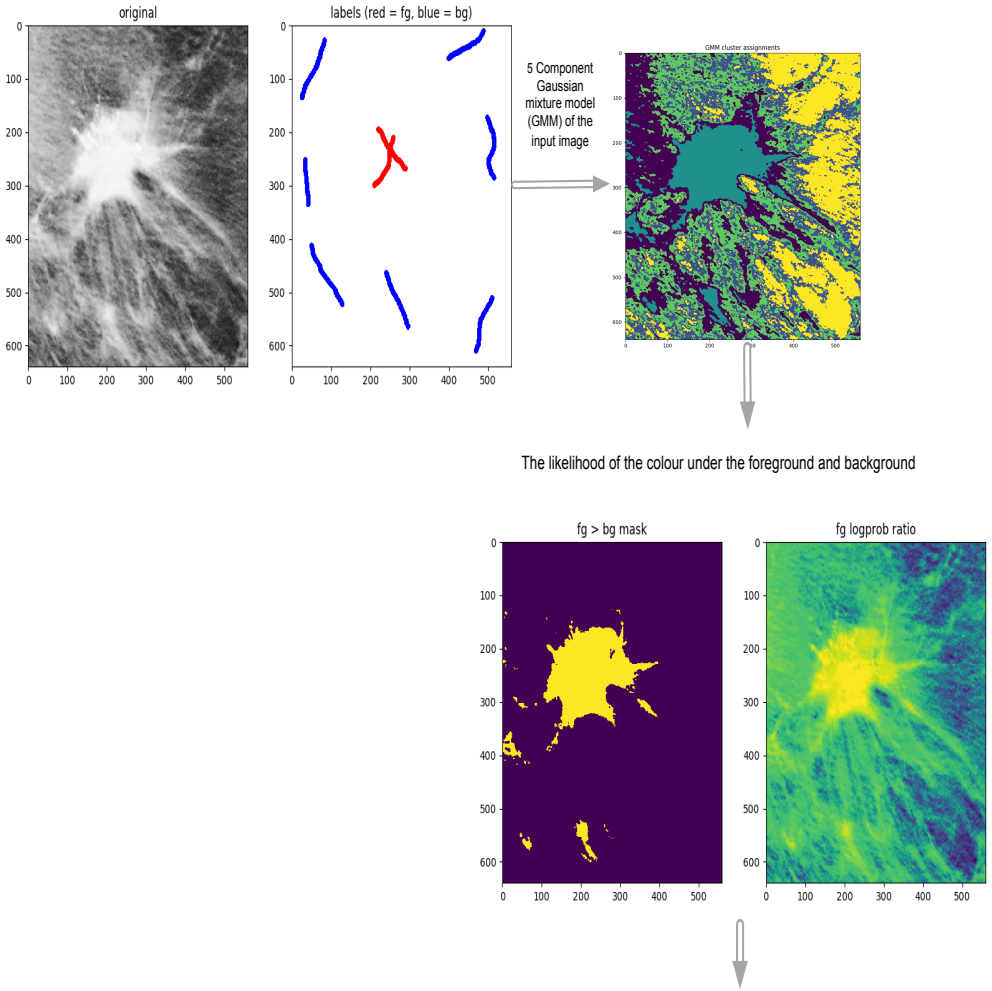


FIGURE 3.8: Schematic overview of the graph-cut segmentation and cooperative cut segmentation.

cancers are interpretable via them. For instance, the breast density, and texture information are essential to analyze breast cancer. In this study, we extract three kinds of features from ROIs: gray level intensity, gradient information, and morphological features. Besides some other features such as fractal dimension features is also obtained from ROIs.

Morphological Features

We adopt a set of features to provide geometric information from the segmented objects. The most basic features are central moments and perimeter. More advanced extracted features are skewness and compactness.

Besides, we provide a group of features on the model rubber band straightening transform called the RBST. These features are extracted from the surrounding mass areas, in which an area along the contour of the segmented object (see Figure 3.9) is chosen and converted by the RBST into a Cartesian coordinate system. The mass boundary area is transformed into a rectangular image. The branches along the spiculated mass are captured in the rectangular transformed image. We take the second order gradient of the RBST image for counting the number of branches. Table 3.1 reports some morphological features with their corresponding brief explanation.

TABLE 3.1: Some extracted morphological and texture features.

Features	Explanation
Feature from an RBST image	The features include short-run emphasis, long-run emphasis, run-length nonuniformity that are extracted in various angles.
Eccentricity	The ratio of the long axis to the short axis on an RBST image.
Perimeter	No. of pixels on the ROI contour.

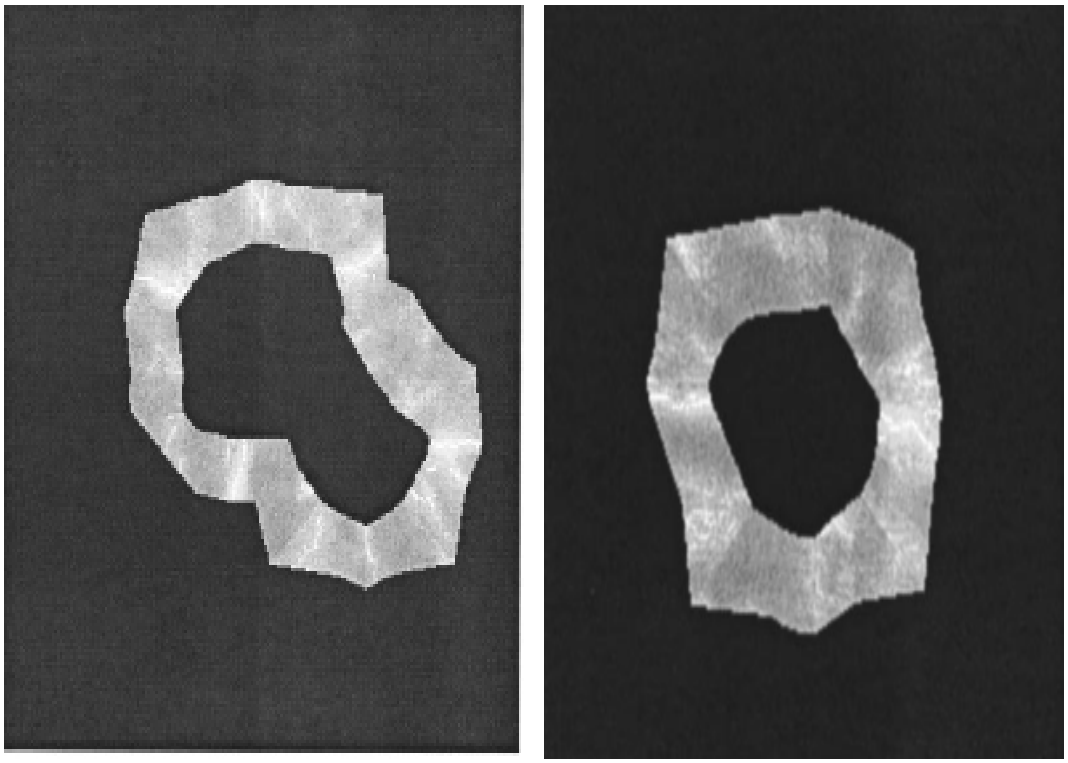


FIGURE 3.9: The boundary mass areas from two samples.

3.2.7 Intensity-based Features

In a spiculated mass area, the average intensity decreases as the radius increases, and the standard deviation increases with increasing the radius. We apply these ideas on the divided ROIs into rings and circular sectors along their angles. First, each extracted ROI is divided into a set of nested rings $[r_1, r_2, ..., r_n]$ along the radius. Then, the mean and standard deviation of the intensity are calculated within each ring $\{m_1^r, ..., m_n^r\}$, and $\{std_1^r, ..., std_n^r\}$. We derive the mean, and the standard deviation vector including the minimum min_m^r, maximum max_m^r, as well as the average mean avg_m^r and standard deviation std_m^r sets of all the rings. Besides, we use a linear regression model to estimate the varying rate of the mean and standard deviation along a radius.

This idea is applied on circular sectors of ROI, in which a ROI is divided into $[s_1, s_2, ..., s_m]$ along the angle, and features are extracted from sectors. We drive the mean, and standard deviation vectors. Then, the minimum min_m^s, min_{std}^s, maximum max_m^s, max_{std}^s, average avg_m^s,

TABLE 3.2: The intensity and gradient-based features which are extracted from divided ROIs into rings, and circular sectors.

Ring based features	Circular sector based features
min^r_m	min^s_m
max^r_m	max^s_m
min^r_{std}	min^s_{std}
max^r_{std}	max^s_{std}
avg^r_m	avg^s_m
avg^r_{std}	avg^s_{std}
std^r_m	std^s_m
std^r_{std}	std^s_{std}

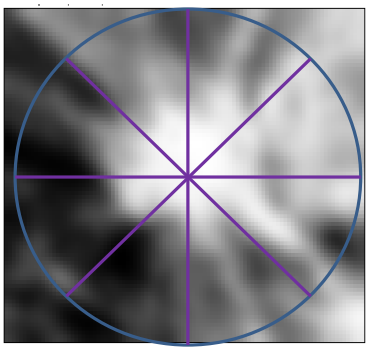
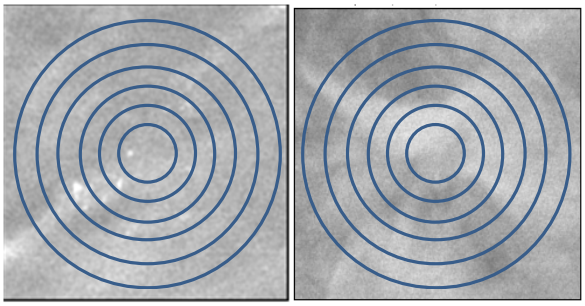


FIGURE 3.10: Extracted sectors along the radius of a 2D sample.



(a) **(b)**

FIGURE 3.11: Extracted rings along the radius of a 2D micro-calcification sample in (a), and a mass sample image in (b).

avg^s_{std}, and standard deviation std^s_m, std^s_{std} are calculated of the mean, and standard deviation vectors.

3.2.8 Gradient-based Features

Again, we use the extracted rings from ROIs to derive gradient features. First, the gradient of ROIs is calculated in each ring, then for each ring, the number of pixels with gradient magnitude above the threshold as well as the direction of gradient parallel to radius are calculated. The percentages of high magnitude pixels are determined for each ring and reported as its gradient feature value. Moreover, we divided each ROIs into sectors and calculated the rate of pixels where the gradient magnitude is above a threshold, and the gradient direction is perpendicular to the radius.

3.2.9 Fractal Dimension Features

Several studies on biomedical frameworks have observed that the analysis of the fractal property of self-similarity provides significant information from the insight of the biomedical system under investigation. Fractal dimension (FD) as a non-integral dimension is dependent on the complexity of the dimensionality of an object, in which it quantifies how the object fills the space. The most common way to estimate FD of a digital image is the representing the characteristic of the image using the function of the scale parameter. The log-log domain of the function is drawn, then using the linear regression the slope of the straight line to the curve is calculated.

We estimate the fractal dimension from ROIs as another feature. The 2D power spectrum is mapped to the radial (f, θ) space from the Cartesian coordinates (u, v) space. It is mapped using the re-sampling by the weighted average of the four neighbors of each point for radial distances ranging from zero to half of the sampling frequency over the angles of $\theta = [0, 179^o]$. Then, the 2D spectrum in the (f, θ) space is transformed into a 1D function $S(f)$. The spectrum $S(f)$ is related to the radial frequency f by the $S(f) \propto (\frac{1}{f})^{\beta}$ formula. The slope β is estimated of the fitted line to the curve in the log-log scale of $S(f)$ using the linear regression using the linear regression approach. The FD is determined as $FD = \frac{8-\beta}{2}$.

Feature Selection

We provided various dimension reduction methods such as principal components analysis (PCA), linear discriminant 2D projections, random trees embedding projection, and Hessian locally linear embedding [94].

A PCA approach is a linear transformation method. The goal of this method is to analyze data for identifying pattern in which it reduces the dimensions of the data with minimal loss of information. In this method, we are interested to find the directions that maximize the variance in the dataset. This approach uses the orthogonal transformation to convert the data correlated variables into the linearly uncorrelated variables. These linearly uncorrelated variables are called principal components. The resulting vectors are not correlated orthogonally, so the transformed data has the lower variance than the original data.

Linear discriminative analysis 2D projection approach likewise a PCA method is a linear transformation technique. This method aims to project a feature space onto a smaller subspace while preserves the class-discriminatory information.

The random trees embedding projection as the space partitioning technique is adapted to the intrinsic dimensionality of the data. It applies the $k - d$ tree data structure to the data that partitions hyper-rectangular cells, the splits along different directions with a binary tree caused the leaves that contain individual cells. This method guarantees the reductions of the cells [28].

Recovering the low-dimensional data from the high dimensional with the assumption that the data lie on the manifold M in which it is the sub-manifold of the Euclidean space. The Hessian locally linear embedding method recovers a parametrization of data that lie on the manifold in which it is locally isometric to the subset of the Euclidean space. The Hessian is defined using the orthogonal coordinates on the tangent planes of the Euclidean space.

At the beginning of our study, we trained different classifiers on our data such as support vector machines (SVM), and decision tree classifiers. For evaluating each classier, we used different feature selection approaches. For example, for the SVM classifier, the random trees method was

used. Figure 3.12 illustrates techniques that are applied to the extracted feature vectors to reduce their dimensions.

For the multiple-instance learning classifier in our framework, we use the linear discriminant 2d projection method [89] as the feature selection approach to select the best features within each type of the extracted features.

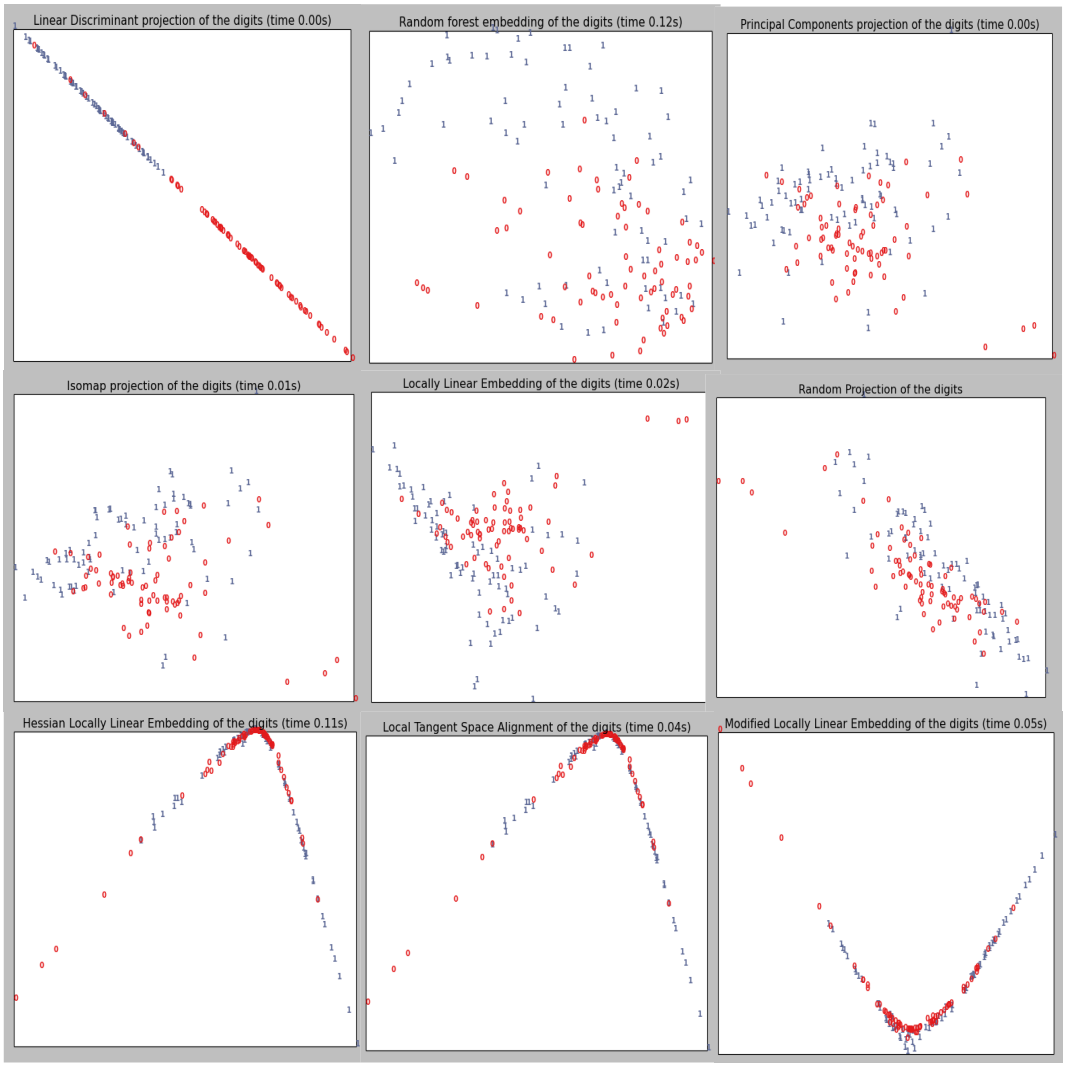


FIGURE 3.12: Different dimensionality reduction approaches.

3.3 Mass Classification

The classification stage aims to classify DBT using all derived features from ROIs. The multiple-instance learning classifier annotates DBT images using the combination of extracted features from all ROIs over the entire DBT. Our goal is to reduce the false positive detection rate and the false negative rates

3.3.1 Multiple-instance Learning

Dietterich [31] introduced a weakly supervised learning framework called multiple instance learning (MIL). In a MIL, we are given a training data which includes annotated bags whereas a bag consists of a set of unlabeled instances [108]. In this model it is assumed that each instance has a label (positive or negative); however, this label is not given in the training stage. A bag is labeled positive if at least one of its instance is positive. In a MIL, we are given a training set consisting of n bags $B_1, B_2, ..., B_n$ instead of individual instances, where each bag $B_i = \{x_{i1}, ..., x_{ik}\}$ consists of a varying number of instances, and it aims to learn instances and bags labels:

$$\mathrm{h}(x_{ij}) : X \longrightarrow [-1, 1] \tag{3.13}$$

where $\mathrm{h}(.)$ represents the instance-level learning, and the bag-level learner is defined:

$$\mathrm{H}(B_{ij}) : X^m \longrightarrow [-1, 1] \tag{3.14}$$

$$\mathrm{H}(B_i) = \max_j \{\mathrm{h}(x_{ij})\} \tag{3.15}$$

A binary MIL model usually uses the standard negative likelihood as the loss function during the learning procedure [98]. The MIL framework has been used in various medical imaging applications to perform medical image classification, learning features and image segmentation [68], [109], [99]. There is also a rich literature of studies on the synergy of classification methods and the MIL setting. These have been applied to medical images including cancer detection techniques such as multiple cluster instance learning in [98] for colon cancer detection.

MIL on DBTs

In the classification stage, our goal is to annotate DBT images using learned features from the previous step. In this study, the multiple-instance learning strategy is used for the DBT classification, which is supposed to mimic the same radiologists analyzing DBTs. A DBT is annotated positive if at least a corresponding slice has an early sign of breast cancer. Consider the training bag is a DBT image, and its 2D slices are the corresponding instances as shown in Figure 3.13. The detail of the learning procedure will be given later in chapter 4.

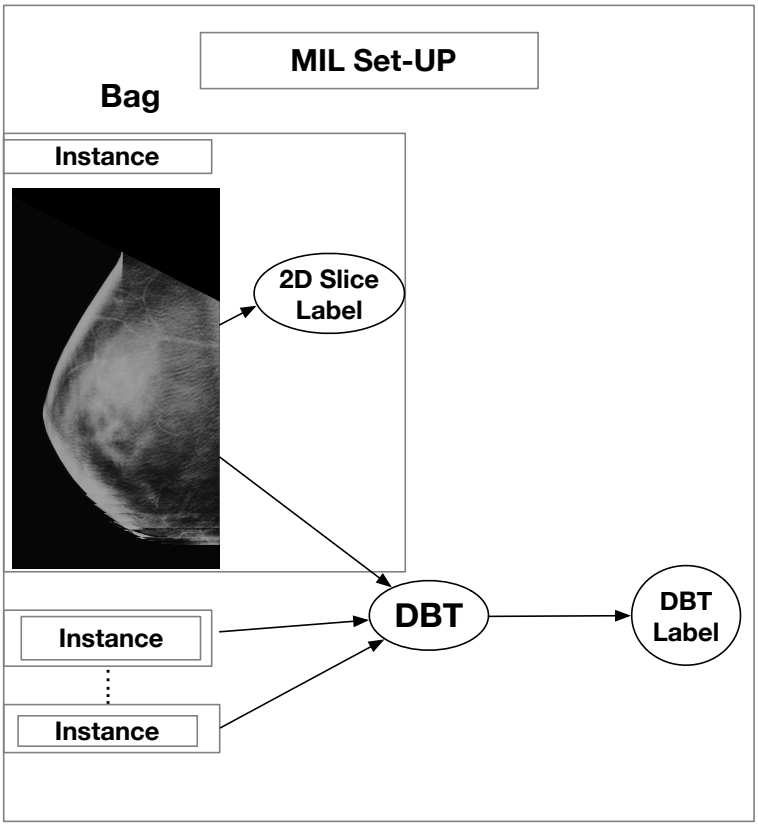


FIGURE 3.13: Schematic overview of the MIL setting on a DBT.

3.4 Micro-calcification Detection

The micro-calcification detection task either by radiologists or CAD frameworks is often a tricky job. Micro-calcifications may be occluded by surrounding breast tissues or may be confused with

other white noises in X-ray images. There are limitations in the visibility of micro-calcifications in DBTs for some factors: large search space, lower signal-to-noise ratio, blurring, and reconstruction using thick slice spacing [78].

In this study, we developed the separate channel in the CAD framework to detect micro-calcifications. However, due to the insufficient micro-calcification data (2 DBT images), this framework serves only as the first step for the future works. In this study, we developed the separate channel in the CAD framework to detect micro-calcifications. However, due to the insufficient micro-calcification data (2 DBT images), this framework serves only as the first step for the future works. Micro-calcifications as an early sign of breast cancer could be visible as a cluster of white dots in an X-ray image. Based on other studies, micro-calcifications absorb more X-ray energy when compared to other breast tissues, in which the intensity range for micro-calcifications is among the top 30% of their surrounding tissues. By utilizing this knowledge that micro-calcifications appear as high-intensity dots in X-ray images, we extracted the foreground from the background. For the foreground extraction, we applied a sliding window, while the window was shifted in an overlapping manner. All pixels with intensity of lower than the threshold of 30% were set to the background (see Figure 3.14). Then we applied the Laplacian of Gaussian filter (LOG) as the kernel for detection calcify dotes on the images. The responses of images were binarized. Moreover, a label connects filter was used for each calcification candidate to label each calcify dot as an object and specify it with identification. The local density and the size of each calcification blob are essential factors for micro-calcifications detection. We applied these constraints to the suspicious calcifications recognized from the filtered images. We also added the appearance of frequency within adjacent 2D slice constraint using the extracted 3D information [103].

3.5 Bilateral Asymmetry Detection in DBT Images

Bilateral asymmetry is the lack of symmetry between medical images of the left and right breasts. The asymmetry in a breast includes asymmetry in shape, surface, and architecture of a breast.

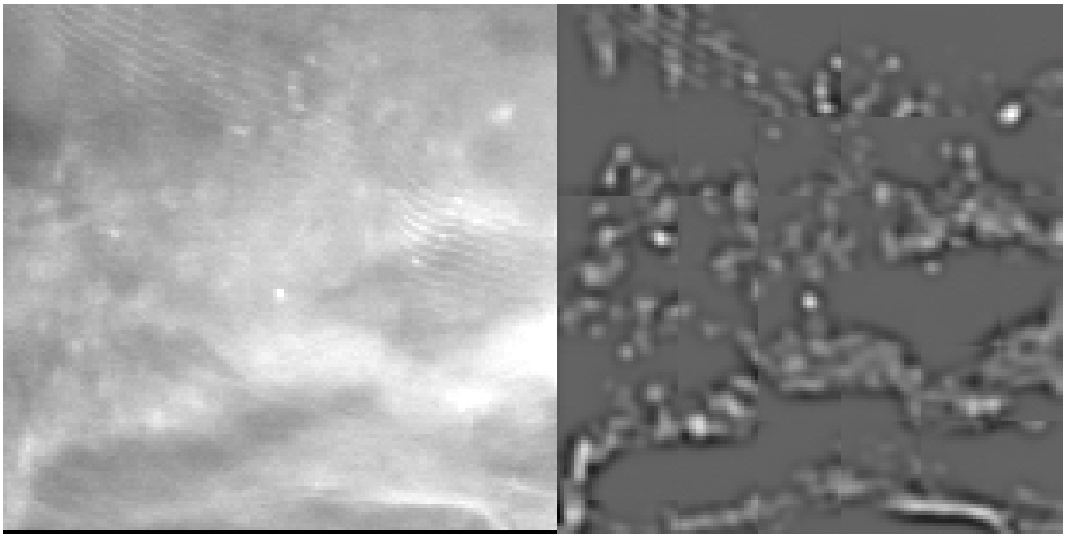


FIGURE 3.14: The image region before and after foreground extraction.

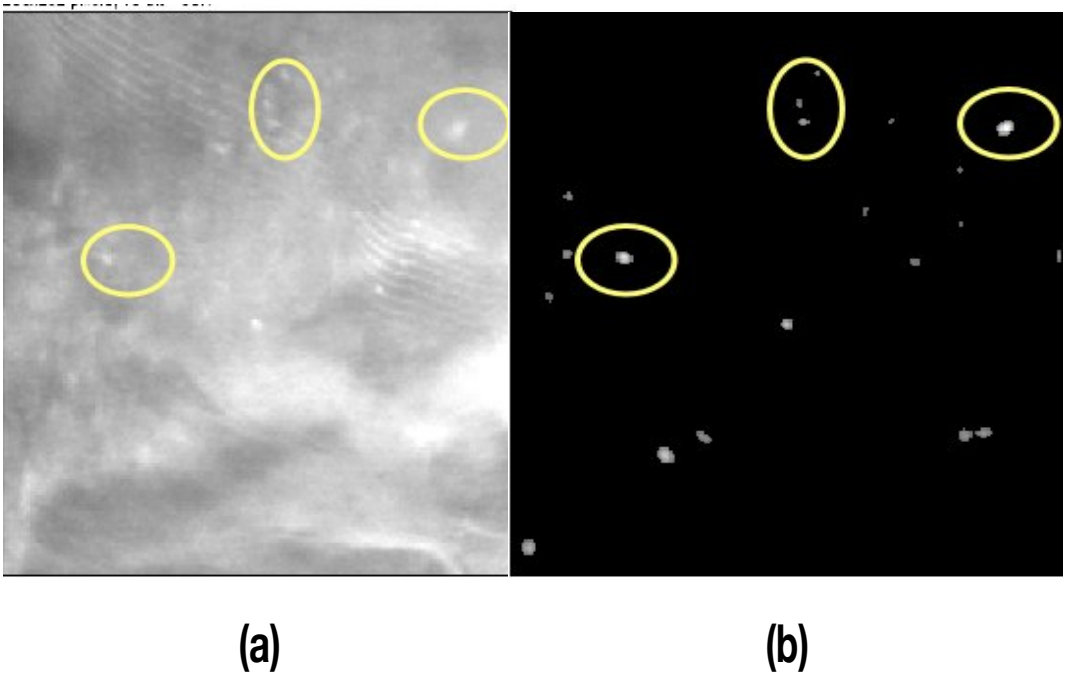


FIGURE 3.15: Colored ovals in (b) represent extracted micro-calcifications from the image in (a).

The two breasts (left and right) are similar in 97% of healthy women. The appearance of asymmetry may be caused by the developing masses or the poor image capturing process.

For the bilateral analysis in DBT, radiologists must compare all slices of both breasts. Thus, a

CAD system for bilateral analysis in DBT can assists radiologists in reducing workload and also allows them to examine the bilateral asymmetry of the breasts quickly, and precisely. The limited study has been done to design these CAD bilateral analysis systems for breast cancer screening on DBT images. The primary challenge in this task is finding the proper breast registration or breast alignment approach.

We introduce the bilateral DBT images asymmetry detection scheme based on analyzing left-to-right 2D slice images from corresponding DBT images. The system comprises three steps (see algorithm 1). In the first step, the proper fiducial (control) points are extracted from the left and the right 2D slices. Then, the registration algorithm is applied to the extracted fiducial points from both slides images. Finally, using the variety of distant metrics, the alignment images are compared. Our dataset contains 60 non-cancerous DBT images, and 27 cancerous cases. The non-cancerous cases were obtained from 30 patients, in which images were taken from left and right breasts. We do not have images with positive bilateral labels in our cancerous cases. So for evaluation of the bilateral asymmetry approach, we can only use the non-cancerous DBT images.

Before selecting the fiducial selection points from 2D images, the image is flipped pixels so that the informative pixels start at the pixel in the $(0, 0)$ coordinate. The integrated intensity value chooses the side of the breast on the upper left corner.

Fiducial Points Extraction

The fiducial points prepare the critical piece of information in the process of image registration on DBT images. In this study, we extract fiducial points that along the breast contour. We detect the contour of the breast in the images with the simple thresholding way of taking advantage of the data that background pixels have zeros. The contour is cleared up with the smoothing process. Then we conduct a sampling step along the smooth contour to collect the fiducial points, and we also obtain a set of points located on the left border of images (see Figure 3.16).

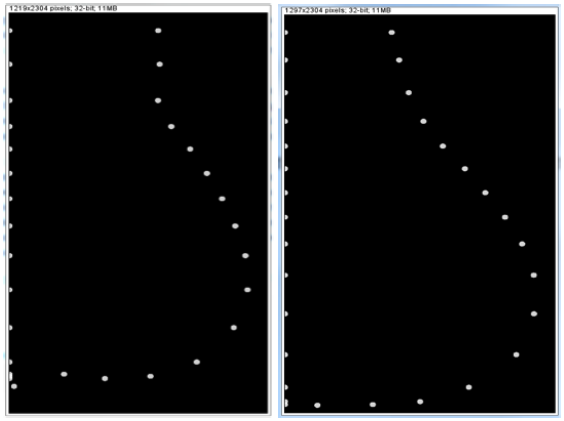


FIGURE 3.16: The selected fiducial points from both sides of the breast images.

The Thin-plate Spline Registration

The breast registration is defined as the spatial mapping between the left and the right breasts. Common developed breast registration methods are elastic body spline, B-spline, and thin-plate spline-based methods [37]. The goal of breast registration is to find the optimal transformation that aligns one image to another one.

The thin-plate spline interpolation called TPS was introduced in [17]. TPS has been used widely in medical imaging as the registration or the deformation method [107]. This name refers to a physical analogy that involves the bending of a thin sheet of metal. Likewise, the metal that has rigidity, the TPS deformation resists the bending by implying the penalty including the smoothness of the fitted surface. In the metal setting, the deflection is in the Z direction that is orthogonal to the plane. This idea has interpreted the lifting of the plate as the displacement of the X or Y coordinates within the plate for the problems of the coordinates transformation. In this study, we use the TPS interpolation method [18] on DBT slices. This registration function has two terms the bending, and the affine term. Given a set of fiducial points, the displacements of all other points in the image are determined by the bending term that defines the fiducial points, and the affine term that establishes the mapping into the infinity:

$$f(x, y) = a_1 + a_x x + a_y y + \sum_{1}^{k} w_i U(|F_i - (x, y)|) \tag{3.16}$$

where $f(x, y)$, and F_i represent the displacement of an arbitrary 2D point (x, y), and the fiducial

point, respectively, and a_1, a_x, a_y, and w_i are spline coefficients. The final summation term controls the bending influenced by k fiducial points using a controlling function $U(r) = r^2 \log(r)$, where r is the Euclidean distance $\sqrt{((x_i - x)^2 + (y_i - y)^2}$ from the ith fiducial point to (x, y). In fact, TPS defines a surface on fiducial points, while the function $f(x, y)$ minimizes the bending energy with $\triangle^2 f = 0$. The solution to the linear system is derived by solving the above equation plugging the source and destination control points. The TPS algorithm results in aligned images from source and destination images (see Figure 3.17).

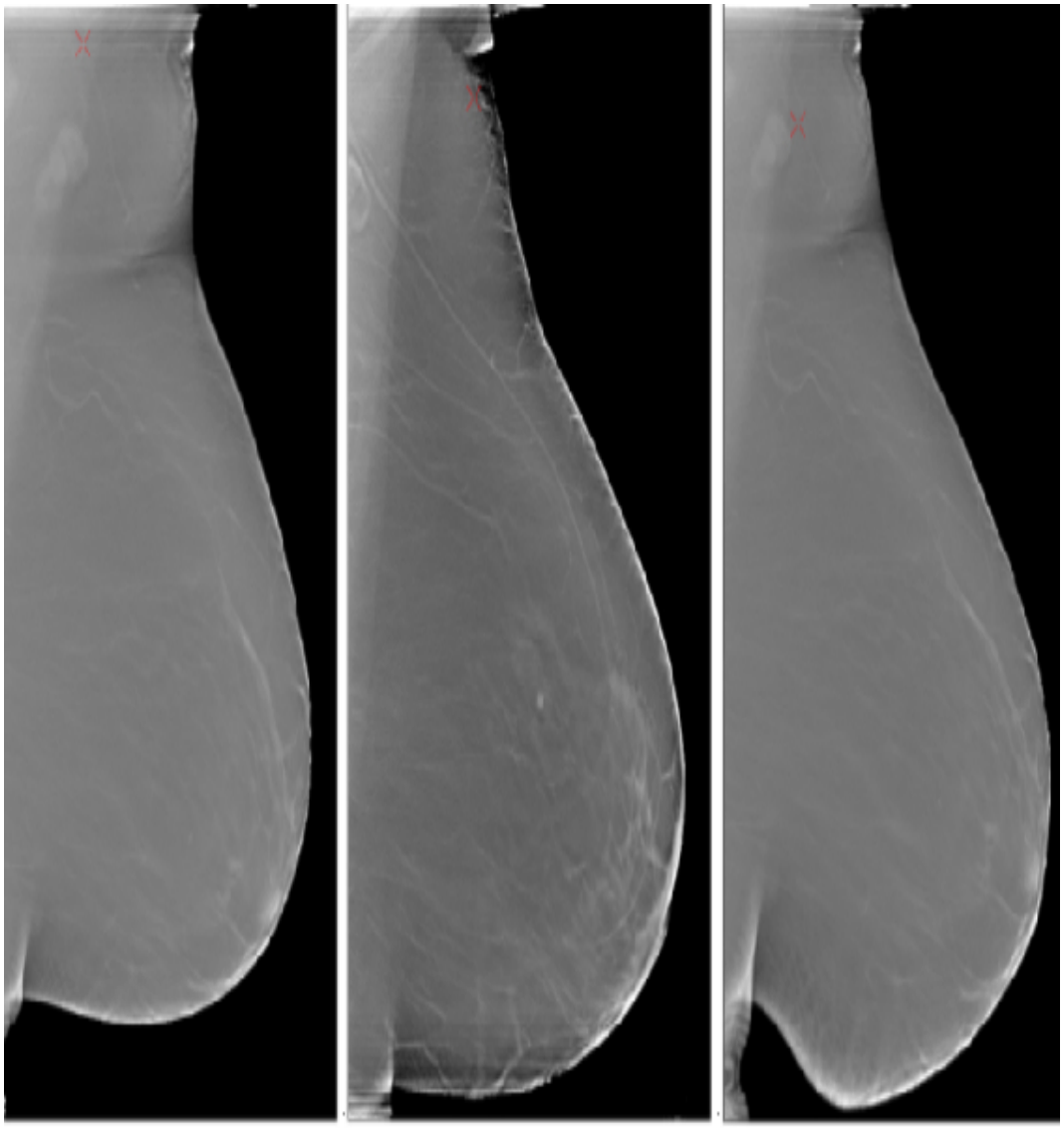


FIGURE 3.17: The horizontally flipped right breast (on the left), the left breast (in the middle), and the aligned right breast (on the right).

Region Comparison

The last step of this procedure is exploring some distance metrics to measure the symmetric between two aligned images. First, we crop the image into the fixed sized regions to create objects for comparison. Note that the areas are cropped with an overlap of half the size so that no boundary information is lost. We use two types of distance metrics in this study, the intensity-, and the feature-based methods. The intensity-based techniques are computed based on intensity values distribution divergence, cross-correlation of templates, etc. We are also interested in geometric characteristic in pixels that in our case are the vascular structures, etc. In the feature-based metrics, we use the representative features, such as principal components decomposition and the nonnegative matrix factorization as the ways to extract principal vectors. Note that, we only use the first few principal components for the comparison.

The image matrices are converted into one-dimensional arrays for comparing between the two image regions based on the intensity metric. Then, we apply a few primary distances, such as Euclidean, Manhattan, and Chebyshev. Also, more advanced distances like Pearson and Chi-Square correlation are used for the comparison between cropped regions. Besides, we compare two matrices of areas, in which we use the normalized SQDIFF, and the normalized cross-correlation techniques:

$$R(x,y) = \frac{\sum_{x',y'}(\mathrm{T}(x',y') - \mathrm{I}(x+x',y+y'))^2}{\sqrt{\sum_{x',y'}\mathrm{T}(x',y')^2 \sum_{x',y'}\mathrm{I}(x'+x,y'+y)^2}} \tag{3.17}$$

$$R(x,y) = \frac{\sum_{x',y'}(\mathrm{T}(x',y')\mathrm{I}(x+x',y+y'))}{\sqrt{\sum_{x',y'}\mathrm{T}(x',y')^2 \sum_{x',y'}\mathrm{I}(x'+x,y'+y)^2}} \tag{3.18}$$

where T and I are regions of the template image (the destination), and the source image, respectively.

Algorithm 1 Bilateral Asymmetry Detection.

Input: DBT images from the left, and the right breasts.
Output: Returns the asymmetry measurement images outputs from the input images.

SET k equals to No. of slices from the left DBT.
SET counter variable equals to zero.
for all Slices belong the left DBT image. **do**
    Find the corresponding slice from the right DBT image.
    Flip pixels of slices so that they appear on the same side of the image.
    Find fiducial points from both 2D slices.
    Apply the TPS registration algorithm to map the slice image from the right image onto the left one.
    for all Measurement algorithms **do**
        Construct the empty result images from slices.
        Crop the regions for images.
        for all Regions in images **do**
            Measure the symmetry between alignment breasts, with the parameter.
            Change the result images pixel values with the measurement results.
        end for
        Save the resulting image for the measurement parameter.
    end for
end for

Chapter 4

Deep-learning based CAD frameworks

Most available CAD methods for detecting breast cancers rely on hand-crafted feature extraction approaches. This dependency has limited these systems to what humans understand as interpretable difference among the early signs of breast cancers and the healthy breasts in images. Concerning noticeable internal and external observer variabilities that may affect these critical image-based breast cancer risk measurements [50], we can imply these measurements cannot be consistent.

Additionally, in the last few years, deep learning models have achieved notable performance in various medical imaging tasks such as image analysis and classifications [44], [22], [97]. The human brain's architecture inspires deep learning models, in which they involved algorithms with the structure of artificial neural networks that are inspired by the function of the brain. Deep refers to the number of layers in the network. Deep learning models can represent an input at many levels of abstraction when compared to the conventional machine learning algorithms with the shallow designs. The deep architecture allows the algorithm to learn features automatically without using hand-crafted features. The ability to determine the model pattern without using the feature designing as the most challenging and time-consuming part of the learning algorithms makes deep learning models extremely attractive in the field of medical imaging.

In this chapter, we introduce two deep learning based CAD frameworks (cardinality restricted Boltzmann machines- and convolutional neural network-based systems) for detecting masses in DBT images. For training these models, we used the enlarged dataset which will be discussed in Chapter 5, The deep learning models in CAD frameworks provide the hierarchical 2D slices

representations. These models comprise of different layers to represent low-to-high level features from 2D slices, in which the layers that are adjacent to the input layer learn more generic, and the layers adjacent to the output layer learn more specific features from images. The learned features from 2D slices undergo the multiple-instance learning algorithm in the CAD frameworks for classifying DBT images. The learned features are the instances, and DBT images are the bags for the multiple-instance learning classifier.

4.1 Methodology

These CAD frameworks include the preprocessing, the deep learning based feature learning, and the multiple-instance learning classification modules (see Figure 4.1). The input images are the 2D slices from DBTs. The preprocessing approach is the same one used in the hand-crafted feature-based CAD system in Chapter 3. The deep learning models learn features from the images through the hierarchical structure automatically. Using the multiple-instance learning (MIL) algorithm DBT images are classified based on the learned features from their corresponding 2D slices. In this thesis, we use the synergy of multiple-instance learning and the random forest (RF) and support vector machines (SVM). For the RF-based MIL, the RF is initialized by training it on instances that inherited their labels from DBT images where they lie. Then at an iteration of training procedure of the synergy of multiple-instance learning and RF, for each tree, a pseudo label for each instance is drawn based on the probabilities that stored in the leaf of trees from the previous iteration. Note that if we have a positive bag (DBT) with no positive instance (2D slice), we assign the positive pseudo-label to the instance with the highest probability of being positive. For SVM-based MIL classifiers, there are several ways to extend the SVM optimization to the MIL one. For example, using the heuristic two-step algorithm in which first the discriminative optimal is found based on the pseudo labels for 2D slices in positive DBTs. Then, the labels for the 2D slices in positive bags are updated based on the optimal discriminative function. This two-step is continued until no more label changes for instances in positive bags. In this study, we used four different SVM-based MIL classifications.

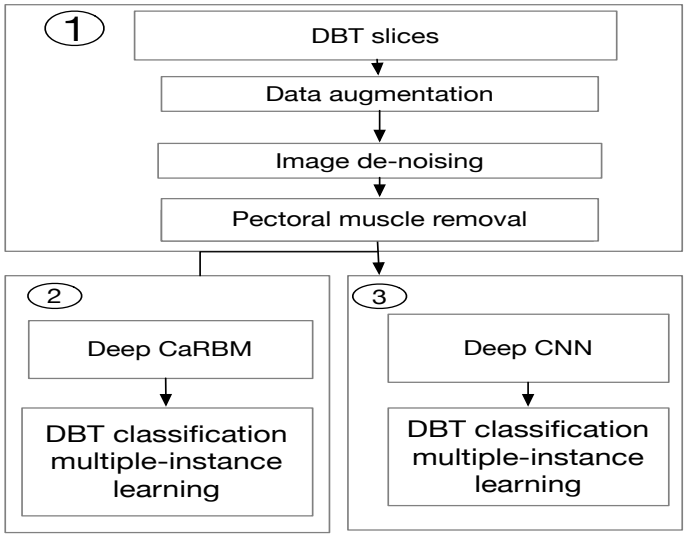


FIGURE 4.1: Flowchart of deep learning based CAD frameworks.

4.2 Deep Belief Networks

This section provides the background information for the deep neural networks. We discuss some prior learning representations in computer vision that are related to deep learning models.

4.2.1 Autoencoder models

Autoencoder is the most straightforward unsupervised neural network (NN) algorithm that can extract features from the data automatically. An autoencoder approach as a type of feed-forward NN that processes an input sample through a multiple-layer neural network to reconstruct the input sample at the output of the network (see Figure4.2). Each layer of the NN comprises a matrix of weights W, biases vectors b, and a form of nonlinearity function such as sigmoid, $\tanh$, or rectified linear unit [105]. The nonlinearity is applied to the input to provide the output of each layer as:

$$\mathrm{h}(\mathrm{v}) = \sigma(W\mathrm{v} + \mathrm{b}) \tag{4.1}$$

where v, and σ are the input and the sigmoid function, but it could be any activation function in an NN. The squared distance between the output of the last layer of the NN (h), and the targets (y)

defines the object to be minimized in the learning procedure. Usually, the stack of autoencoders is used as the deep autoencoder model. For training the stack of autoencoders, first all layers are trained, then the entire NN is fine-tuned for reconstructing the input [105].

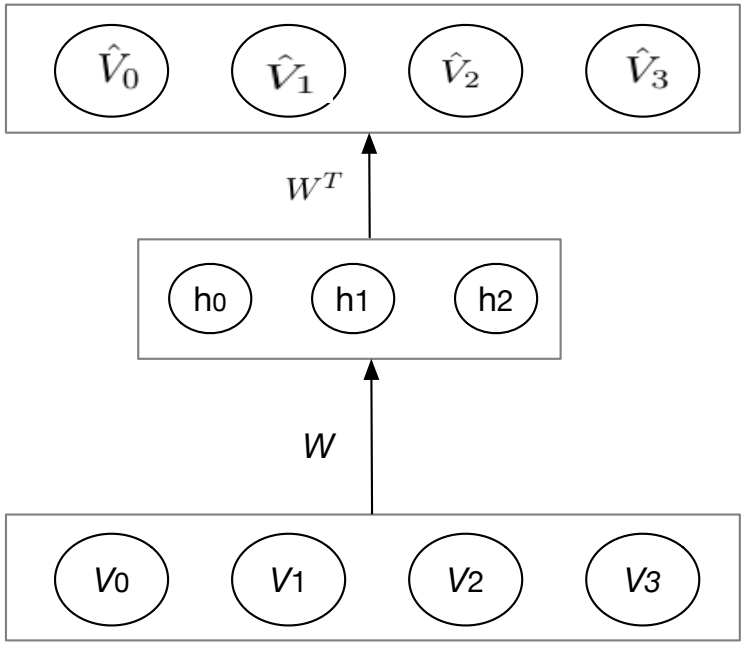


FIGURE 4.2: The autoencoder takes the input v in the first layer, encodes it through the weights W to get a set of feature responses which are mapped through the nonlinearity to get activations in the hidden layer (middle layer). In the output layer (the last layer) the reconstruction of the input as $\hat{\mathrm{v}}$ is provided.

4.2.2 Restricted Boltzmann Machine

A restricted Boltzmann machine (RBM) is an autoencoder model that was introduced in [40]. As shown in Figure 4.3, an RBM is a particular type of Markov random field (MRF) with a hidden layer and a visible one. Nodes of an RBM are connected to each other across layer nodes, but no nodes that are in the same layer. The RBM is trained by minimizing the following energy function:

$$\mathrm{E}(\mathrm{v},\mathrm{h}) = -\mathrm{v}^T b^{\mathrm{v}} - \mathrm{h}^T b^{\mathrm{h}} - \mathrm{v}^T W \mathrm{h} \tag{4.2}$$

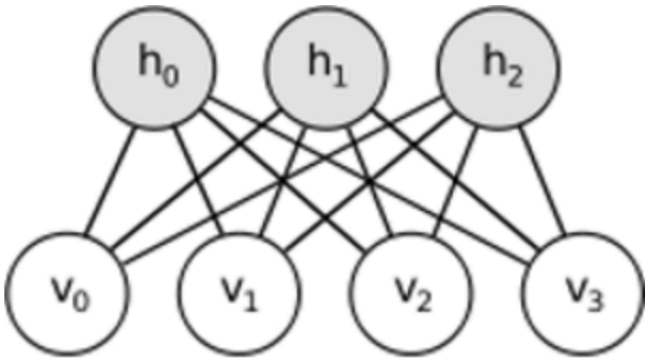


FIGURE 4.3: An RBM with four visible, and three hidden units.

where $\mathrm{v} \in \{0,1\}^{N_\mathrm{v}}, \mathrm{h} \in \{0,1\}^{N_\mathrm{h}}$, and $\theta = \{\mathrm{b}^\mathrm{v} \in \Re^{N_\mathrm{v}}, \mathrm{b}^\mathrm{h} \in \Re^{N_\mathrm{h}}, W \in \Re^{N_\mathrm{v} \times N_\mathrm{h}}\}$ are the bias vectors for the visible (input), and the hidden layer and the visible-to-hidden weight matrix, respectively. The parameters of an RBM are learned by maximizing the probability of the joint configuration $\{\mathrm{v}, \mathrm{h}\}$ that is defined based on the energy function as:

$$\mathrm{P}(\mathrm{v}, \mathrm{h}) = \frac{1}{\mathrm{Z}} \cdot \exp(\mathrm{v}^T \cdot W \cdot \mathrm{h} + \mathrm{v}^T \cdot b^\mathrm{v} + \mathrm{h}^T \cdot \mathrm{b}^h) \tag{4.3}$$

$$\mathrm{P}(\mathrm{v}; \theta) = \frac{1}{\mathrm{Z}(\theta)} \cdot \exp(-\mathrm{E}(\mathrm{v}; \theta)) \tag{4.4}$$

where $\mathrm{Z}(\theta) = \Sigma_\mathrm{v} e^{-\mathrm{E}(\mathrm{v};\theta)}$ is the normalizing constant. Using the maximum-likelihood (ML) method, the parameters of an RBM are learned. The parameters of the RBM θ with the sample data $\mathrm{V} = \{\mathrm{v}_1, ..., \mathrm{v}_N\}$ are computed by the gradient ascent [23]:

$$\theta^{t+1} = \theta^t + \eta \frac{\partial L(\theta; \mathrm{V})}{\partial \theta} \tag{4.5}$$

where η is the learning rate. The L is the average likelihood:

$$L(\theta; \mathrm{V}) = \frac{1}{N} \Sigma_{n=1}^{N} log P(\mathrm{v}_n; \theta) = -\langle \mathrm{E}(\mathrm{v}; \theta) \rangle_0 - log(\mathrm{Z}(\theta)) \tag{4.6}$$

where $\langle . \rangle_0$ is the average of the empirical distribution $P_0(\mathrm{v}) = \frac{1}{N}\Sigma_{n=1}^{N}\delta(\mathrm{v} - \mathrm{v}_n)$. Then we should compute the:

$$\frac{\delta L(\mathrm{V};\theta)}{\delta\theta} = -\langle\frac{\delta \mathrm{E}(\mathrm{V};\theta)}{\delta\theta}\rangle_0 + \langle\frac{\delta \mathrm{E}(\mathrm{V};\theta)}{\delta\theta}\rangle_\infty \tag{4.7}$$

where $\langle . \rangle_\infty$ is the average with respecting to the model distribution $\mathrm{P}_\infty(\mathrm{v};\theta) = \mathrm{P}(\mathrm{v};\theta)$. In e.q 4.7, the first average is computable using the sample data V, but the second average includes the $\mathrm{Z}(\theta)$ partition function that we do not know the value of this partition function so it cannot be computed efficiently 4.7. The standard method for approximation of the average over the distribution is the Markov chain Monte Carlo approach. But this way takes a long time since it requires the estimations of averages that have the expectational amount of terms, and also it produces the large variance of the estimation of the gradient. Hinton [40] introduced the contrastive divergence (CD) algorithm as the approximate ML learning algorithm. The ML method minimizes Kullback-Leibler divergence:

$$KL(P_0||P_\infty) = \Sigma_{\mathrm{v}} P_0(\mathrm{v}) log(\frac{P_0(\mathrm{v})}{P(\mathrm{v};\theta)}) \tag{4.8}$$

The CD learning follows the estimated gradient of the objective function that is the difference between two Kullback-Leibler divergences:

$$CD_n = KL(P_0||P_\infty) - KL(P_n||P_\infty) \tag{4.9}$$

In the CD learning, the Markov chain is started at the data distribution, and the chain runs only for a few steps. In comparison to the Markov chain Monte Carlo approach, in the CD method, the computation and the variance of the estimation of the gradient are reduced per each gradient step. Moreover, the CD learning yields good estimates for parameters, and it has been applied successfully to various type of random fields. In training an RBM usually, the CD algorithm is used for the maximum likelihood.

The inference in an RBM is the single feed-forward encoding (see e.q 4.10):

$$\mathrm{h}(\mathrm{v}) = \sigma(W\mathrm{v} + b^{\mathrm{h}}) \tag{4.10}$$

where

$$\sigma(x) = \frac{1}{1 + exp(-x)}$$

σ is a logistic function. Selecting the sigmoid non-linearity can be justified based on that it is equal to the expectation of the conditional distribution $P(\mathrm{h}|\mathrm{v})$ [86]:

$$\sigma(W^T\mathrm{v} + b^{\mathrm{h}}) = \mathrm{E}_{P(\mathrm{h}|\mathrm{v})}[\mathrm{h}] \tag{4.11}$$

Typically a deep RBMs (DRBMs) model comprises a stack of RBMs. For training a DRBMs, the stack is initialized using the greedy layer-wise pretraining process. We enter an RBM in the empty stack and train it until a good representation of the input is obtained. Then, another RBM is added on top of the first one, in which it takes the feed-forward activations from the hidden layer of the lower RBM as its input. This procedure is continued for more RBMs in the stack. Then, the stack of trained RBMs is converted into a feed-forward NN, and the entire NN is fine-tuned with backpropagation technique.

The backpropagation is a way of computing gradients of expression using recursive application of the chain rule. In an NN, gates get some inputs that can compute its outputs values and the local gradient of data. These gates do these computational independently in the forward pass. Once the forward pass is over, during the backpropagation, gates can learn about the gradient of its output on the final layer of the entire network. Each gate takes this gradient and multiplies it to all the local gradients for all its inputs. So choosing the proper non-linearity function for activation neurons is the critical task in designing the NN since the gradient of the non-linearity as the local gradient is multiplied to the gradient of this gate's output for the whole objective. The backpropagation identifies as the way for gates to communicate with each other in the network, so that they make the final output value higher.

Hinton [74] introduced another way to design a DRBMs. This model comprises a stack of RBMs that are connected in the bidirectional fashion. The bidirectional connections between RBMs help the model to be trained jointly instead of greedy layer-wise training.

4.3 Deep Cardinality Restricted Boltzmann Machines MIL CAD Framework

Conventional CAD frameworks such as the proposed CAD framework in chapter 3 heavily rely on the hand-crafted human features. Even though this approach is successful on the available data; the classifier can be fooled by the feature extraction or other preprocessing stages failures with additional clinical cases. So a method which relies on less prior processing such as deep learning models is desirable. In this section, we introduce the deep cardinality restricted Boltzmann machine MIL CAD framework for detecting masses in DBT. As shown in Figure 4.5, the DCaRBM-based CAD framework comprises the input, the DCaRBM, and the MIL classification. The DCaRBM learns image patterns from 2D slices through a deep learning fashion, and the MIL classifies DBT images based on the learned features from 2D slices.

4.3.1 Deep Cardinality Restricted Boltzmann Machine

An RBM often is used as a pre-training method in the deep belief networks (DBNs) models. Using DRBMs, the performance of various machine learning tasks in different domains have been promoted. However, the disconnection between the unsupervised nature of an RBM and its final task where the learned features are used such as the classification task may decline the task-specific performance of an RBM. One of the ways to improve the task-specific performance of an RBM is to incorporate sparsity into the learned representation of the RBM [86]. A sparse coding model aims to learn the over-complete set of basis functions for representing the input data. So methods with a variety of tasks that use the sparse representations can achieve the promising results.

Sparse Coding

In the sparse coding model, an over-complete set of basis functions are learned [64]. Then using the sparse distribution of coefficients or feature activations for each basis function, a small number of the basis functions reconstruct the input data. For the learning procedure in a sparse coding model, the cost function is given by:

$$\mathrm{C} = \frac{\lambda}{2}\|\mathrm{Wz\text{-}x}\|_2^2 + \|\mathrm{z}\|_1 \tag{4.12}$$

where W, z, and x are the matrix of the basis functions, the feature activation, and the input vector, respectively. The W is not a square matrix since it shows the over-complete basis functions. As shown in e.q 4.12, the cost function comprises of two parts: the reconstruction term and the weight matrix. The reconstruction part enforces the input reconstructed from the learned features. The weight matrix of the cost function imposes the inferred feature maps to be sparse in the ℓ_1-norm [90]. The aim of the sparse coding is minimizing the cost function over the entire dataset. For optimizing the cost function, the gradient descent is employed on the W that is shared between all input values to update it. Then, z is inferred for each input based on the current value of the W.

The sparsity can provide the interesting features that are invariant to local transformation. However, with increasing the sparseness of the solution, the discriminative information is limited to the few activations. So the sparsity usually is used as the part of the learning model. For example, the combination of the hierarchical models with sparse coding that proved the robust classifications with the ability to learn the different levels of features from the entire image [16].

DCaRBM as the Sparse Model for DRBM

We can apply sparsity by a direct sparse model or by an undirected way. The direct sparse models impose the sparsity using a target sparsity penalty for activations and the sparse model that separates activations into spike and slab units [36]. While in the undirected sparsity ways, the sparsity is applied to the model by the hard sparsity constraints directly on the energy function.

Authors in [88] proposed the method based on assigning the performance to counts over the subset of binary variables called the cardinality potentials. The cardinality potentials are useful for applying the undirected sparsity to approaches.

We use the sparse deep restricted Boltzmann machine (RBM) model called deep cardinality RBM (CaRBM) [86]. This model uses the undirected way to apply the sparsity into the RBM. Hard constraints are applied to the energy function of RBM using the cardinality potential function. The sparsity constraint is assigned to the energy function by restricting the number of hidden units that decode the visible layer data. The probability of the joint configuration $\{\mathrm{v}, \mathrm{h}\}$ in CaRBM [86] is defined as:

$$\mathrm{P}(\mathrm{v}, \mathrm{h}) = \frac{1}{\mathrm{Z}} \cdot \exp(\mathrm{v}^T \cdot W \cdot \mathrm{h} + \mathrm{v}^T \cdot \mathrm{b}^{\mathrm{v}} + \mathrm{h}^T \cdot \mathrm{b}^{\mathrm{h}} \cdot \psi_k(\sum_{j=1}^{N_{\mathrm{h}}} \mathrm{h}_j)), \tag{4.13}$$

$$\text{s.t.} \quad \psi_k(c) = 1 \quad \text{if} \quad c \leq k, \quad \text{otherwise} \quad \psi_k(c) = 0$$

Function $\psi_k(\sum_{j=1}^{N_{\mathrm{h}}} \mathrm{h}_j)$ in e.q (4.13) is the cardinality potential function with a constant parameter k. The cardinality potential function implements the competition between hidden layer units. This competition provides the sparsity in the model.

The deep CaRBM is the stack of CaRBMs that are considered as pretraining models for the DBN model. The contrastive divergence algorithm is used as the learning procedure in the deep CaRBMs. After several CaRBMs are trained in a greedy-layer wise manner, the network is converted to the feed-forward NN. Note, the non-linearity function in the stack of CaRBM is not the sigmoid function. The non-linearity $\mu(.)$ [86] is used in the stack of CaRBMs that satisfies the equation:

$$\mu(W^T \mathrm{v} + \mathrm{b}^{\mathrm{h}}) = \mathbf{E}_{P(\mathrm{h}|\mathrm{v})}(\mathrm{h}), \tag{4.14}$$

where $P(\mathrm{h}|\mathrm{v})$ is derived from e.q 4.13. Authors in [86] proved that this non-linearity exists and they proposed the alternative method of multiplying by the Jacobian of $\mu(.)$ to backpropagate through $\mu(.)$. In Figure 4.4 we demonstrate the reconstructions of sample input data from the

first layer features maps of DCaRBM.

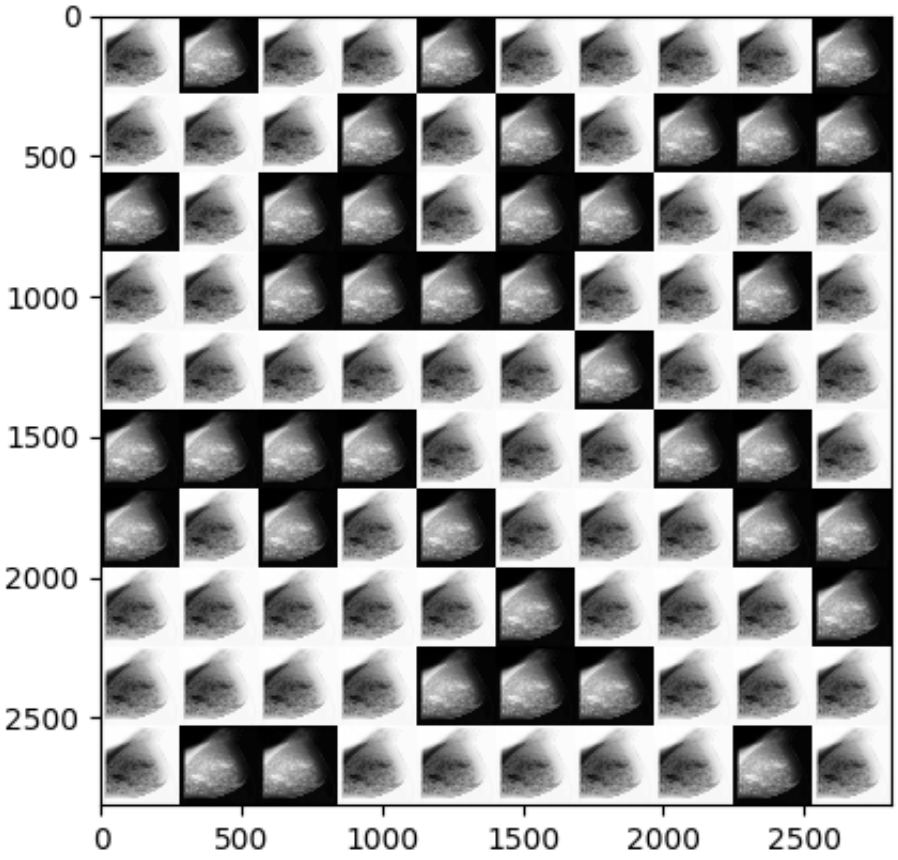


FIGURE 4.4: Image reconstructions in the 100 randomly selected hidden units from the layer one of the DCaRBM on 2D slices samples.

Training the DCaRBM CAD Framework

The input data for the CAD framework includes the original data and the synthetic data that will be discussed in Chapter 5. The preprocessing module is the same one in the pipeline of hand-crafted feature-based CAD framework. The learning procedure for the DCaRBM-CAD system consists of the two step-learning procedure. First, it used the DCaRBMs to learn features from 2D slices. Then using the multiple-instance random forests combines the learned information to classify DBT.

Here in this thesis, a stack of the three CaRBMs is trained in the greedy-layer wise fashion way. Learning the global features via a small number of hidden units in the initial learning epochs that causes other hidden units do not participate in the learning procedure yields the dead units issue. The efficient way to address this problem is augmenting the log-likelihood with the Kullback Leibler (KL) penalty (the negative of the KL divergence in e.q 4.15). The dead hidden units problem appears in the initial learning epoch in the DCaRBM when parameters are initialized randomly. We apply the KL penalty only during the unsupervised learning (training the stack of DCaRBMs), following the model in [86]. The target sparsity is fixed to 10%, the strength of the KL penalty is found so that the model achieves the average hidden unit population sparsity without dead units. The following penalty term is applied to the logarithm object 4.13:

$$\lambda(p \log q_j + (1-p)\log(1-q_j)) \tag{4.15}$$

$$\text{s.t.} \quad q_j = \frac{1}{N}\sum_n P(\mathrm{h}_j = 1|\mathrm{v}_n)$$

where p, λ, N, and n are the desired target expected sparsity, the strength of the penalty, the normalization value, and the number of the visible units in the visible layer, respectively. For training the DCaRM, the input 2D slices are divided into the mini-batch size of 100 samples. The number of hidden nodes from the $1{,}000$ nodes that can explain the visible vector data is fixed to 10 nodes. The stack of DCaRBMs is converted to the feed-forward neural network.

The MIL classifier is trained on the learned features from 2D slice (the hidden units of the last CaRBM). The learning procedures for the MIL methods which are common between three-CAD frameworks will be discussed later in this chapter.

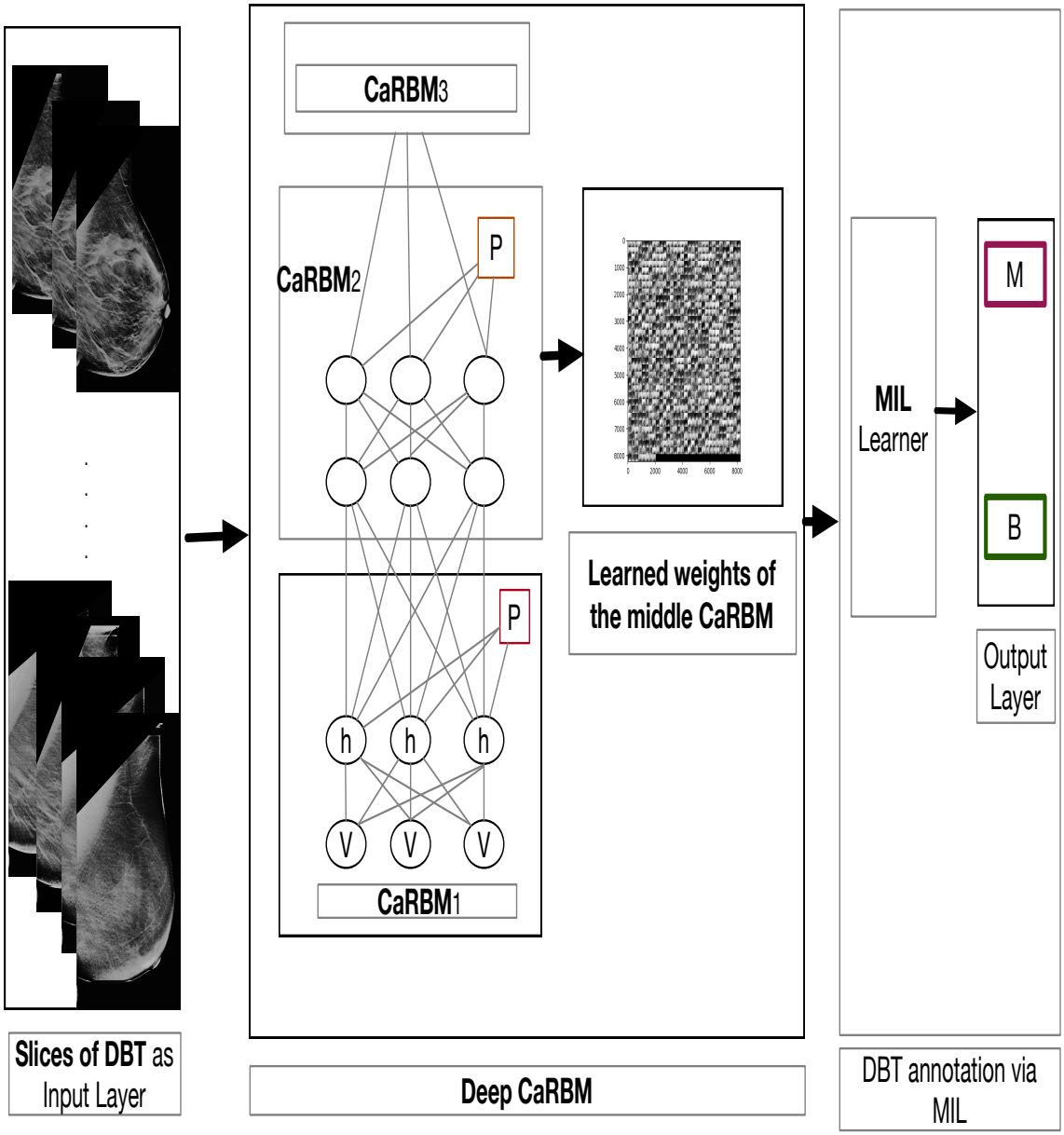


FIGURE 4.5: Schematic overview of the DCaRBM-MIL framework, where V; h; and P represent the visible unit, the hidden unit, and the cardinality potential function in each CaRBM block, respectively. The output layer shows the prediction classes for inputs from M (spiculated mass) and B (benign) classes.

4.4 Convolutional Neural Networks Based CAD Framework on DBT

In this section, we introduce the third CAD framework for detecting masses in DBT images. This framework same as the DCaRBM MIL CAD system comprises three components: the input, the deep convolutional neural network, and the multiple-instance learning. The input data and the

multiple-instance learning sections are like those in the DCaRBM MIL CAD framework, but features learning section in the third CAD model is the deep convolutional neural networks.

4.4.1 Convolutional Neural Networks

Images display plenty of spatial relationships between neighboring pixels. A convolutional network (CNN) can keep these spatial relationships between neighboring images pixels that should not be affected by their location within the image. A CNN incorporates this characteristic of image exhibition into the model by learning a set of N filters $\{f_1, .., f_N\}$. These filters are convolved with the input data to provide 2D features maps [105]:

$$\mathrm{z}_k = f_k \otimes I \tag{4.16}$$

$$f_k \in \{f_1, .., f_N\}$$

where $\otimes$, f_k, and I are the convolution operator, kth filter, and the input image, respectively. The convolution operator is applied to the two inputs and results in the output like one of the inputs that are smoothed or modulated by the other one. This operation provides the output pixels in which they are the weighted sum of neighboring input pixels. If you are familiar with the correlation operation, we like to mention that the convolution is similar to the correlation, but one of the inputs is reversed. For instance, the correlation in 2D with a filter is the convolution with the filter that is flipped up or down and left or right. For clarification of the convolution operation, we expand the e.q 4.16 on the corresponding matrix of the image I as:

$$\mathrm{z}_k = (f_k \otimes I)(x, y) = \sum_{i=-N_{frows}}^{N_{frows}} \sum_{j=N_{-fcols}}^{fcols} f_k(i, j) I(x - i, y - j) \tag{4.17}$$

where (x, y) is the coordinate of the image pixel, and $frows$, and $fcols$ represent number of rows and columns of the f_k, respectively.

4.4.2 CNNs Architecture

CNNs are similar to other NNs: they are comprised of neurons with learn-able weights and biases. Neurons are arranged in layers of the network, but unlike a regular NN, the neurons in a CNN are organized in 3D fashion (width, height, and depth). The primary layers to design a CNN are convolutional, pooling, and fully connected layers (see Figure 4.6). To form a CNN, one or a few alternative convolutions and rectified linear units (ReLU) layers that are followed by pooling layers are stacked, and this pattern is repeated until the input dimensionally to the small size. These layers are followed by the fully connected layers leading into a softmax layer that holds the output such as class scores for the classification task.

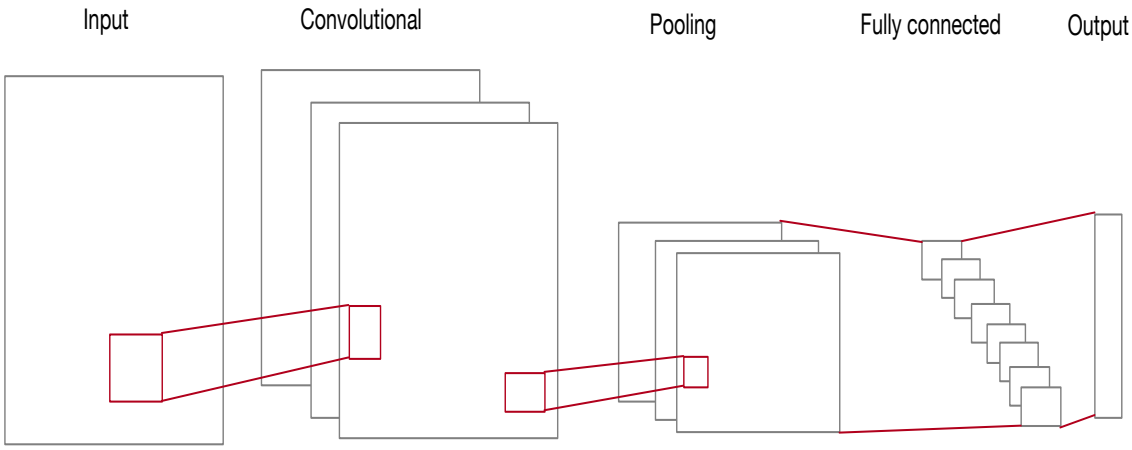


FIGURE 4.6: A CNN model with a convolutional, a pooling, and a fully connected layer.

Convolutional Layer and Its Spacial Arrangement

The convolutional layer computes the output of the neurons that are connected to the local areas in the input image. The parameters in a convolutional layer are a set of the learn-able filters. Filters in convolutional layers are typically small along width and height that extend through the input volume. Each filter is convolved across the width and the hight of the input and computes the dot production between the filter and input at any position. The response of each filter at

every position is the 2D activation map. The stack of these 2D activation maps along the depth dimension results in the output volume.

Neurons in each layer of CNN models are not connected to all neurons in the previous volume; instead, each neuron is connected only to the local region of the input volume. The receptive filed hyper-parameter is defined as the local region in the input space along width and height that is affected by the particular feature of a CNN model. Note that this local connection is only along the width and height but not along the depth of the input volume. For this spacial arrangement, three other hyper-parameters are determined: the depth, stride, and zero padding. Depth value is the number of filters that are used in the same section of the input. The amount of stride is the number of pixels that we slide the filter when applying the convolution. Zero-padding value is the number of zero pads that are added to the border of the image to control the size of the output. The spatial extent of the output volume is defined as below:

$$\text{Output width} = \frac{W - F + 2P}{S} + 1 \tag{4.18}$$

$$\text{Output height} = \frac{H - F + 2P}{S} + 1 \tag{4.19}$$

where W and H represent width and height values of the input, and F, S, and P are the size of the receptive field, the stride value, and the amount of the zero padding, respectively (see Figure 4.7). Note that depth value of the output is equal to the number of the filters are used in the convolutional layer.

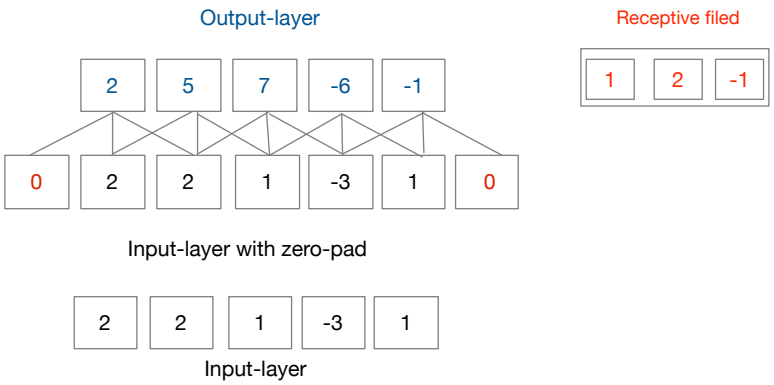


FIGURE 4.7: The receptive filed with the size of $F = 3$, the input size $W = 5$, $S = 1$, and $P = 1$, resulting in the output of size 5.

Pooling and Fully connected Layers

The responses of the convolutional are passed into a non-linearity activation function to produce activation maps h. Typical non-linearity functions that are used in CNN architectures are: sigmoid, tanh, and rectified linear units (ReLU). The activation function is followed by the pooling layer that is responsible for providing invariance to slightly different in the input image. The pooling layer operates down-sampling along the spatial dimension of the volume input. The common pooling operations are average, and max. The former takes the mean of the elements, while the max pooling selects the largest element in each region:

$$\mathrm{p}\ \zeta = \frac{1}{|\zeta|} \sum_{i \in \zeta} \mathrm{h}_i \tag{4.20}$$

$$\mathrm{p}\zeta = \max_{i \in \zeta} \mathrm{h}_i \tag{4.21}$$

where ζ, and i are the small spatial region of each map of activations and indexes the activations in the pooling region, respectively. After convolutional and pooling layers, a CNNs model has one or more fully connected layers. For the classification task, the softmax classifier follows the last fully connected layer:

$$\hat{y}i = \mathrm{softmax}(h_i^l) = \frac{e^{h_i^l}}{\sum_{j=1}^{N} e^{h_j^l}} \tag{4.22}$$

The entire network is trained with backpropagation technique of a loss function such as cross-entropy [105]:

$$\mathrm{L}\ = -\sum_{c=1}^{N} y_c log(\hat{y}_c) \tag{4.23}$$

where $\hat{y}_c$, y_c, and N are the activations of the $l-$th layer of the model h^l in e.q 4.22, the target label for the input image, and the number of classes, respectively.

Advantage of CNN Models

A CNNs model has several advantages:

- The input consists of images; this assumption helps us to encode specific properties in the architecture.
- It is a shift and translational invariance model.
- Only some layers of a CNN contain learn-able parameters such as convolutional and fully connected layers. The activation layers and pooling layers implement fixed functions. However, the pooling, convolutional, and fully-connected may have additional hyper-parameters. For instance, the receptive field is the hyper-parameter in a convolutional layer.
- Each filter is replicated on the entire visual input, in which these units share the weight and bias values. This property of CNN model is called the weight sharing. The assumption that if one feature works well at some spatial positions on the image, the feature may be useful in other areas of the image yields the weight sharing. This property reduces the number of parameters for a CNNs architecture dramatically.
- The extension autoencoder models to the convolutional ones provided models with the ability to encode the shift invariance into the model. Deep RBM was extended to the convolutional model, using the 2D convolutions and the probabilistic form of the max-pooling. This deep model capable of learning curves, and circles from pixels directly, and discriminative features using it on top of scale-invariant feature transform vectors.

4.4.3 Other CNN Architectures

In the previous paragraphs, we described a CNNs architecture with convolutional, pooling, ReLU, fully connected, and softmax layers. Typically, a CNNs architecture utilizes one or multiple of these layers as well as other layers such as cross-channel normalization, and dropout layers. For example, the CNNs in Figure 4.8 includes two convolutional, two pooling, one dropout,

and two fully connected layers. This model uses the adaptive moment estimation (Adam) optimization algorithm [48] for updating the network weights in the training procedure. We will explain the Adam algorithm in section 5.4.3. Figure 4.9 shows another CNNs architecture with the variety of layers as convolutional, max-pooling, fully connected layers as well as normalization, dropout, and deconvolutional layers.

Let us mention the most popular CNNs architectures. LeNe that was used to read zip codes [54]. AlexNet was applied to the ImageNet data [51], ZFNet that was the improvement on AlexNet [106]. GoogLeNet that was the development of the inception module yielded the reduction of parameters [87]. VGGNet that showed the depth of network is important to improve the performance of the model [83]. ResNet skipped the heavy usage of batch normalization technique and also missed fully connected at the end of the network [38].

4.4.4 Deep Convolutional Neural Network-MIL CAD Framework

In recent years, CNNs models have gained the population in the various domains of computer vision [27], [51]. In addition to, a CNNs has made great advancements in medical image recognition such as the analysis of DBT x-ray images [85], [76], [35], MRI images [66], and physiological signals [3].

In this study, we developed the CAD framework that uses the CNNs as part of the system for extraction features automatically from 2D slices of DBT. Compared to other deep learning models, using the CNNs gives us three main advantages of our dataset. First, it achieves rotation and transformation invariance through the max pooling process. In the sub-sampling layer, the maximum is obtained over the neighboring pixels, so the one-pixel shifts cannot change the output of the sub-sampling layer [53]. Since extracted mass candidates are not always centered, this is critical in mass detection processes. Second, the spiculated masses respond strongly to Gabor filters. Third, each neuron in a DCNN connects to a local region of the previous layer output, so the CNN has fewer learn-able parameters when compared with other deep NNs. Together these three factors have yielded dramatic improvements regarding the performance of the CAD framework for detecting masses in DBT.

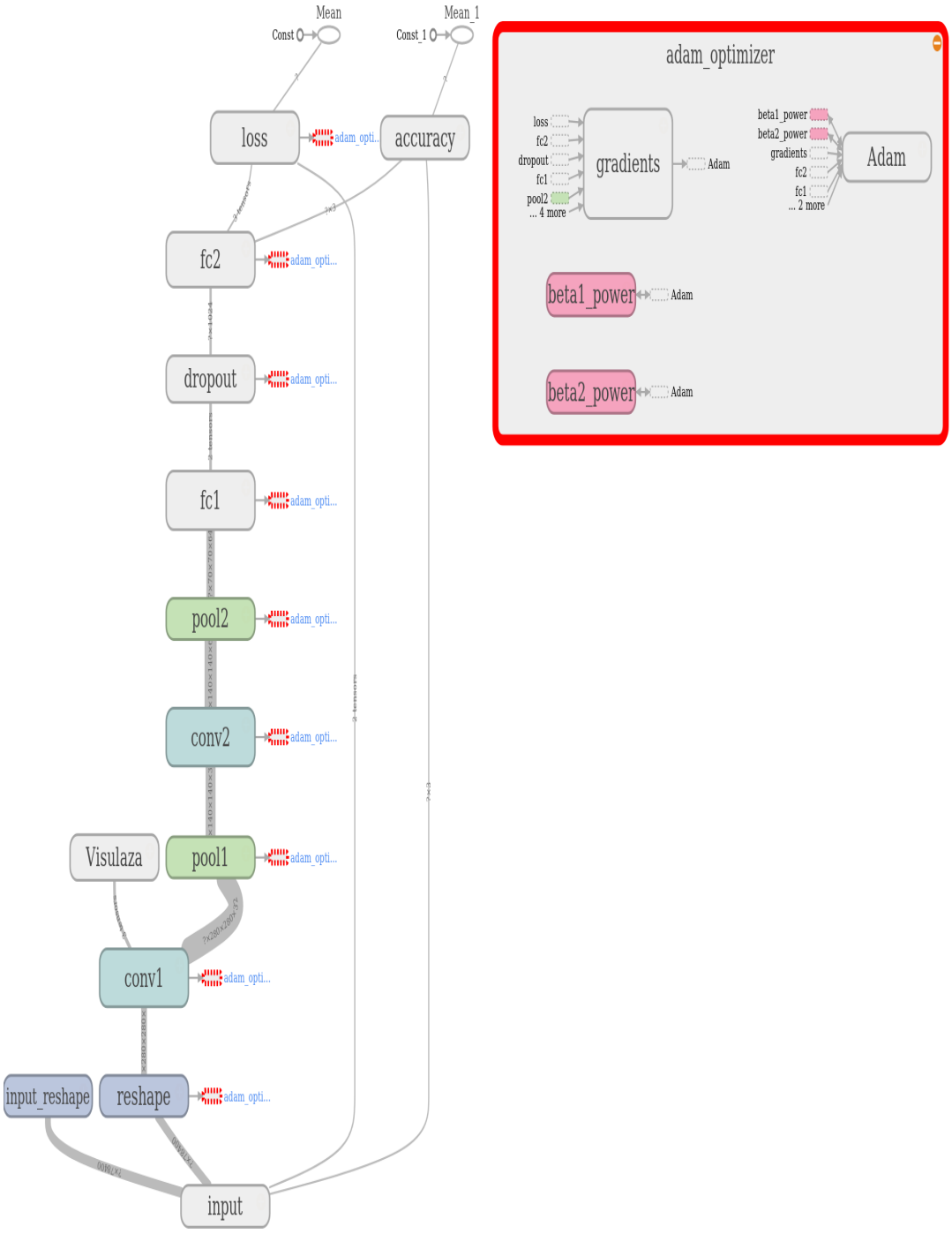


FIGURE 4.8: The CNNs with two convolutional, two pooling, two fully connected, and a dropout layers. This model utilizes the Adam optimizer for optimizing the gradient descent. We used the visualization tool called TensorBoard [2] to draw the CNN Tensorflow graph [2].

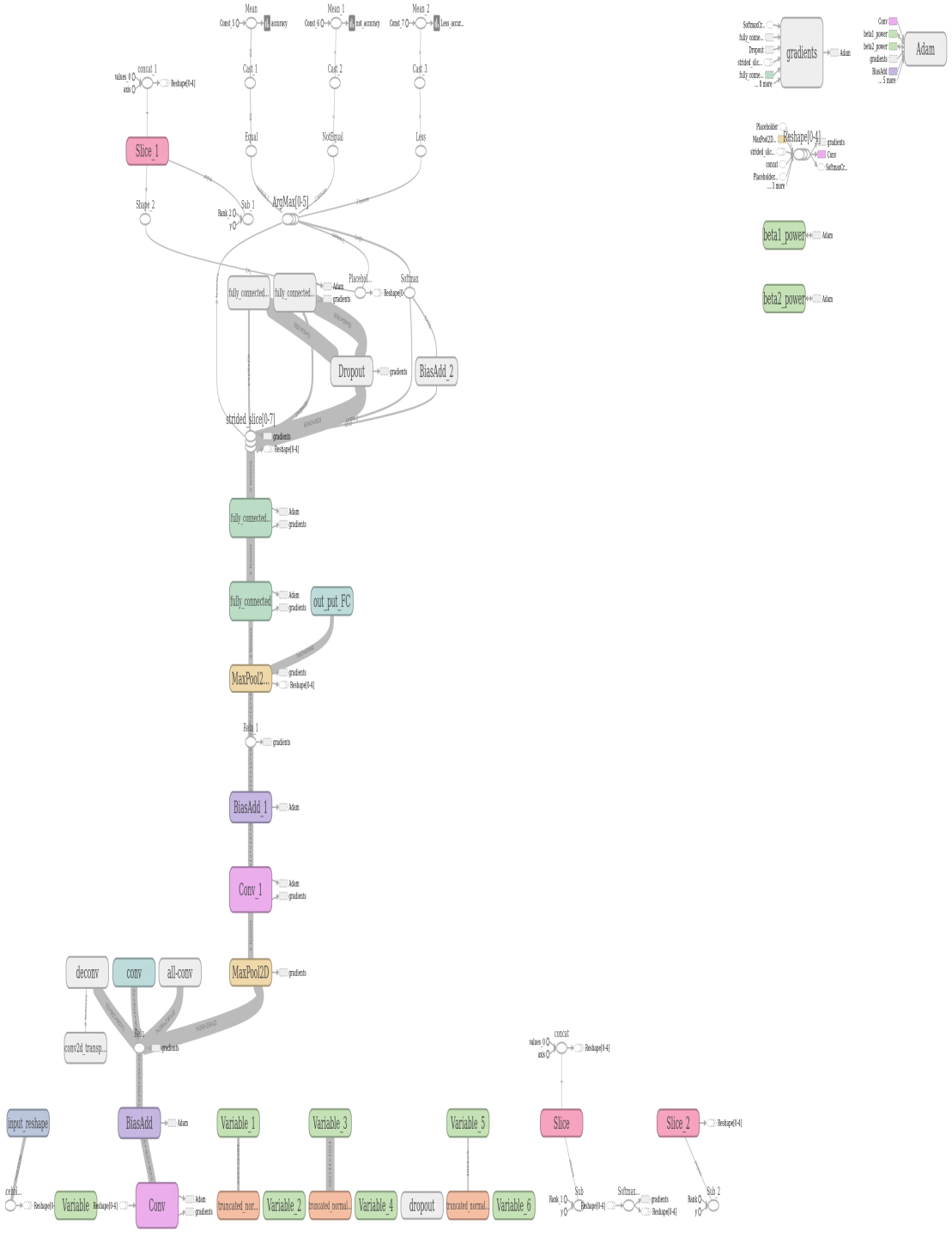


FIGURE 4.9: The graph of the CNN model.

We used the CNNs in the third CAD framework as the feature paradigm (see Figure 4.10) to extract high-level feature representation from the input (256×256 resized image). The CNNs

contains three convolutional and two fully-connected (FC) layers with the rectifier linear unit (ReLU), max-pooling, and batch normalization layers. Three convolutional layers have 16, 32 and 64 filter kernels sized 5×5. Two fully-connected layers contain $1{,}024$ and 2 neurons. All convolutional layers are followed by a ReLU activation function, a max-pooling layer, and a normalization layer. The max-pooling layer has kernels sized 3×3 and is overlapped with a stride of 2 pixels in the max-pooling process. Each feature map is connected to the activation function by the ReLU $\mathrm{f}(x) = \max(x, 0)$.

As shown in Figure 4.10, the network includes batch normalization (BN) layer [42] right after convolutional layer and before the ReLU. The BN technique allows the network to avoid the activation distributions changing issue that is layer-by-layer input distribution variation in the training process of deep learning models. Furthermore, BN facilitates the DCNN training process. The BN transformation normalizes the result by the batch standard deviation, $\sqrt{\sigma_B + \epsilon}$, where the ϵ is the small constant. First, the x_i value is transformed to the normalized one given by:

$$\hat{x}_i = BN_{initilize}(x_i) = \frac{x_i - \mu_B}{\sqrt{\sigma_B + \epsilon}} \tag{4.24}$$

where μ_B is the mean of the input batch. The $BN_{initilize}$ transformation in 4.24 can restrict the input to the activation function to a normal distribution so that it can limit the power of the layer for representation. To address this problem, authors [42] multiplied the provided $\hat{x}_i$ by the scale γ, and shifted by the β:

$$BN_(x_i) = \gamma(\hat{x}_i) + \beta \tag{4.25}$$

Dropout with probability 0.5 that means 50% of the elements are randomly dropped out in the training procedure follows the first FC layer. The softmax layer follows the last FC layer to provide predicted likelihood over labels for each input image. The prediction loss of the network is measured by cross-entropy between the softmax output and ground truth distributions. Figure 4.10 illustrates the order of layers in the network.

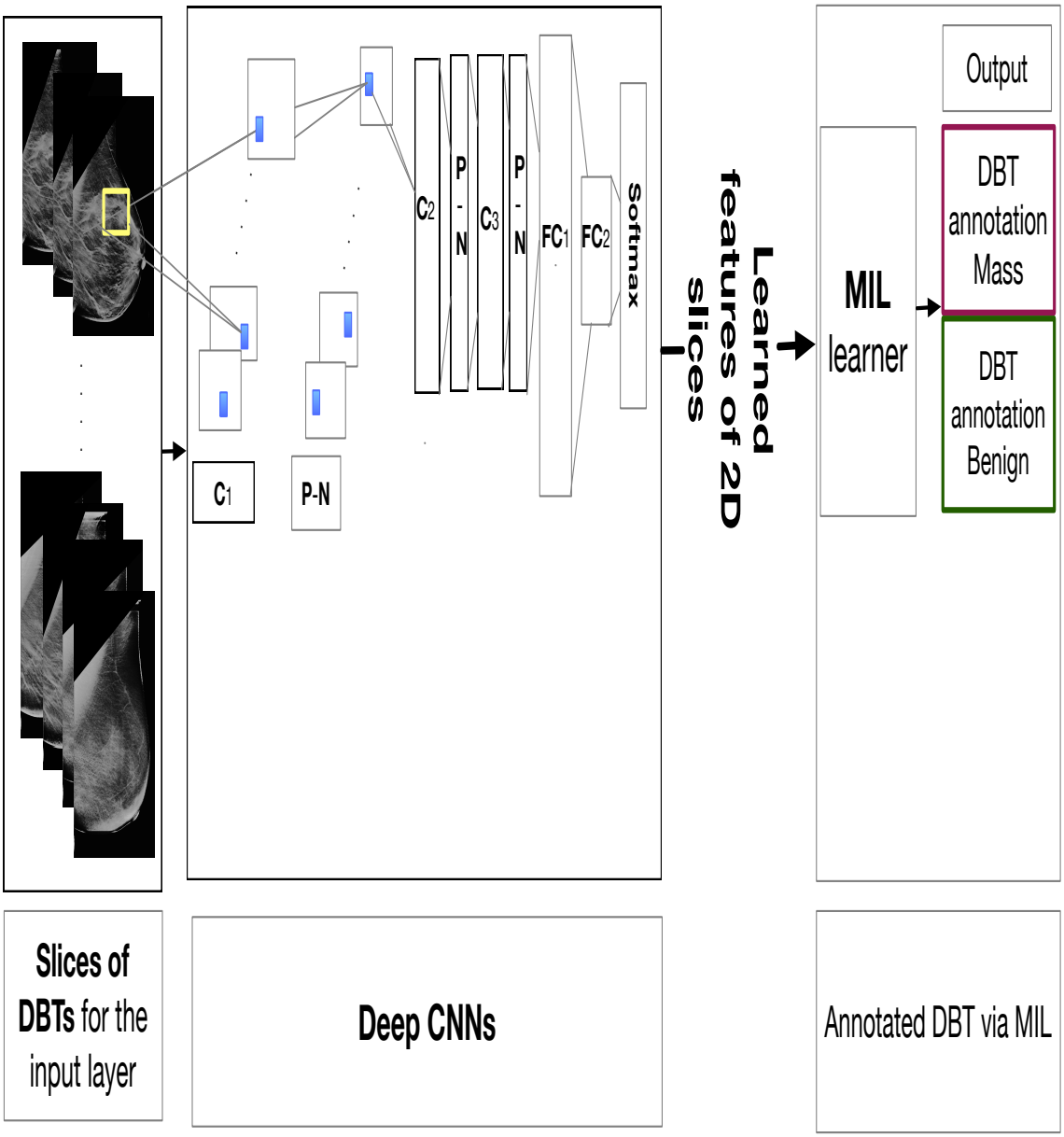


FIGURE 4.10: Schematic overview of the deep CNN-MIL CAD framework, where C, P, N, and FC represent the convolution, the max-pooling, batch normalization, and the fully-connected layers, respectively.

Details of Learning DCNN Parameters

The backpropagation algorithm is used for the training procedure for a CNNs similar to other NNs. We utilize the Adam optimizer algorithm [48] to minimize the cross-entropy loss function. The Adam optimizer works based on the adaptive estimate of lower-order moments, and

it requires only first-order gradient. This algorithm computes the adaptive learning rates for parameters of the loss function. Here, we initialize Adam optimizer algorithm with $\alpha = 0.001$ (the learning rate), $\beta_1 = 0.9$, $\beta_2 = 0.999$ (decay rates for the first- and second-moment estimators) and $\epsilon = e^{-4}$ (small constant for numerical stability). The gradients of the stochastic objective function $f(\theta)$ concerning the model parameters θ at the time t is given by:

$$g_t = \nabla_\theta f_t(\theta_{t-1}) \tag{4.26}$$

The algorithm updates biased first-, and second-moment estimations using the past and past squared gradients:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{4.27}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{4.28}$$

where m_t, v_t are the estimations of the first-moment, and the second-moment of the gradients, respectively. Then the bias-corrected first- and second-moment are estimated:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{4.29}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{4.30}$$

Afterward, the objective function parameters are updated:

$$\theta_{t+1} = \theta_t - \alpha \times \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{4.31}$$

where θ_t is the parameter of the objective function at tth iteration. This procedure continues until the parameters are converged. The Adam optimizer algorithm is applied on 63 mini-batches of 2D sample slices. These samples are extracted randomly at each training iteration. Convolutional layers weights are initialized using Gaussian kernels with a mean of 0 and a standard deviation of 0.1. The biases are initialized to 0 for all layers. DCNN is trained for $5{,}000$ iterations. Training takes around one day on an NVIDIA GTX 1080 11 GB GPU. The trained DCNN learns image patterns through the kernels. The generality and levels of abstraction of the learned patterns

differ across layers of the DCNN model, in which high-level features of the input are provided via FC layers (see Figure 4.11). The outputs of the first FC layer is used as learned features from 2D slices. In DCNN models, the layers adjacent to the input image represent more generic and, the layers adjacent to the output layer provide the more specific level of the input representations.

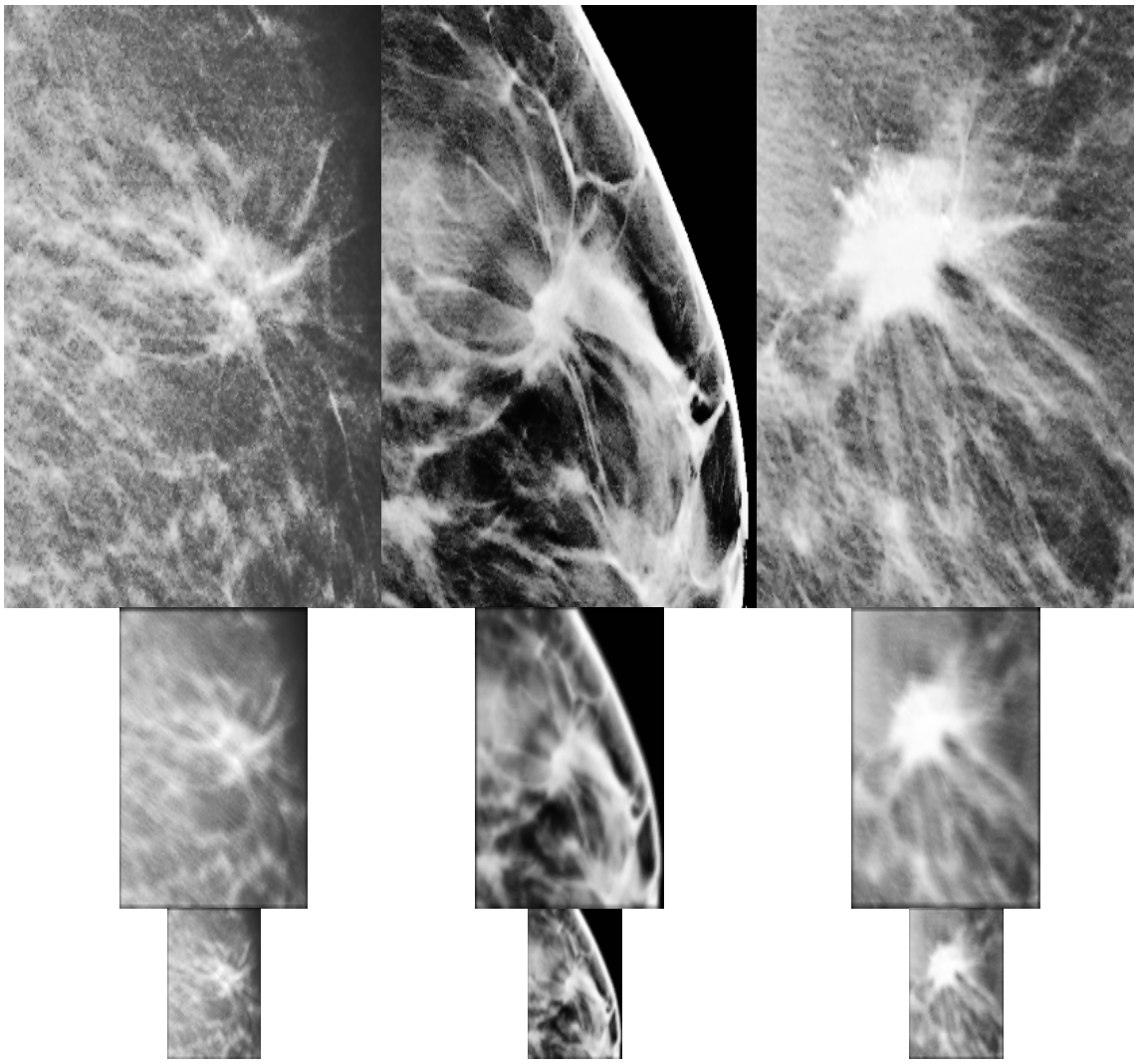


FIGURE 4.11: Hierarchical feature visualization of two slices via DCNNs. Each column shows the slices from top to bottom, middle-level features and high-level features extracted.

4.4.5 Decision Process Using Multiple-instance Learning

After learning features of 2D slices from DBT images, the most challenging task of a CAD framework through slice-based DBT analysis is how to define the integrated classifier to annotated DBT images using the information derived from slices.

Multiple-instance learning (MIL), as a weakly supervised learning algorithm, has achieved impressive performances when annotating data (bag) consisting of unlabeled instances [108]. As mentioned before in Chapter 3 in a MIL model, a bag label is positive if it contains at least a positive instance (see Figure 4.13). This assumption is analogous to the work by radiologists in which a DBT is labeled as positive if at least one corresponding slice image has an early sign of breast cancer (see Figure 4.12). If DBT images are indicated bags and corresponding 2D slices their instances, then the MIL framework is well suited to the task of DBT annotation based on the extracted information from 2D slices.

A MIL learner algorithm is a variant of supervised learning that the available labels for data are ambiguous. In the supervised learning problem, a set of labeled samples are available, and the goal is to learn a classifier using the labeled data. In a MIL problem, the input data are collections of instances (bags). The labels for the bags are available but not for the instances. The goal of the binary MIL classifier is to learn the target that is determined across all positive bags and lost in negative bags. Previously in Chapter 3 we introduced the general representation of the binary MIL classification. In this study, the MIL with multiple classes is desirable because there are more than one early sign of breast cancers (masses, micro-calcifications, bilateral asymmetry, and architecture distortion) that should be recognized in DBT images. The synergy of MIL with other classification methods may provide us the MIL with multi-label classification approach that it will be especially valuable when more clinical cases are used.

There have been various studies on the synergy of classification methods and the MIL setting. One of the significant application of these models on medical images is cancer detection such as multiple cluster instance learning for colon cancer detection [98], mass detection on mammogram images using MIL [33], [29], and using deep multi-instance networks [109]. In this

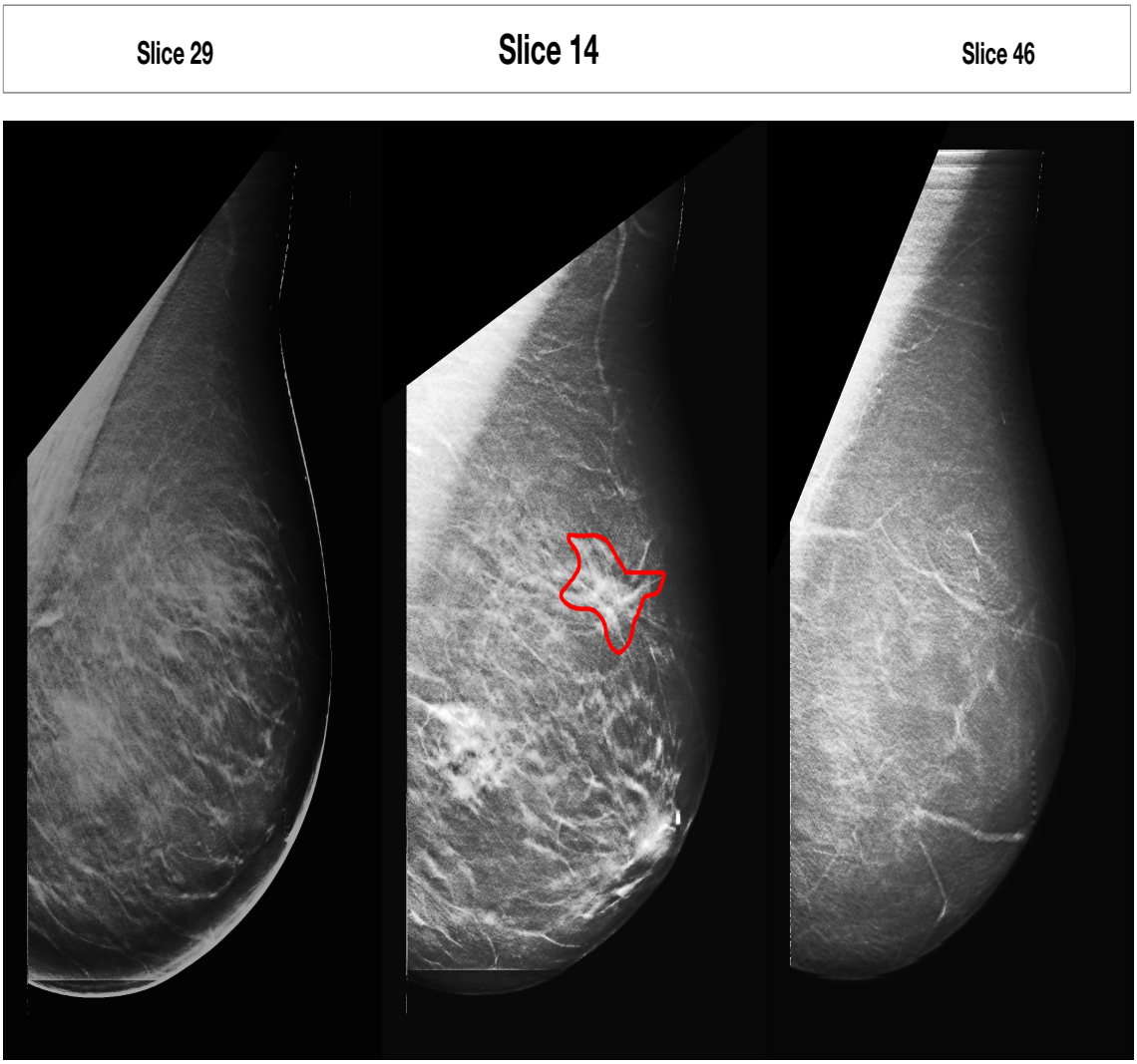


FIGURE 4.12: Three images extracted from the DBT (the 29, 14, and 46th slices). In the middle image, the manual segmentation from 14th slice is in red. DBT image was labeled positive because of the positive (instance) label on slice 14.

thesis, we use the MIL random forests classification which is called MI-RF in [55]. The DBT annotation blocks in Figures 4.5, and 4.10 use the MI-RF to classify DBT images.

Choosing the random forests (RFs) as the synergy classification into the MIL architecture is motivated by several factors. RFs approaches are the comparable state-of-the-art performance in mammograms classification and clustering [13]. Additional, an RFs classier is an inherent multi-class. Besides this classifier provides an easy training procedure for CAD systems [55].

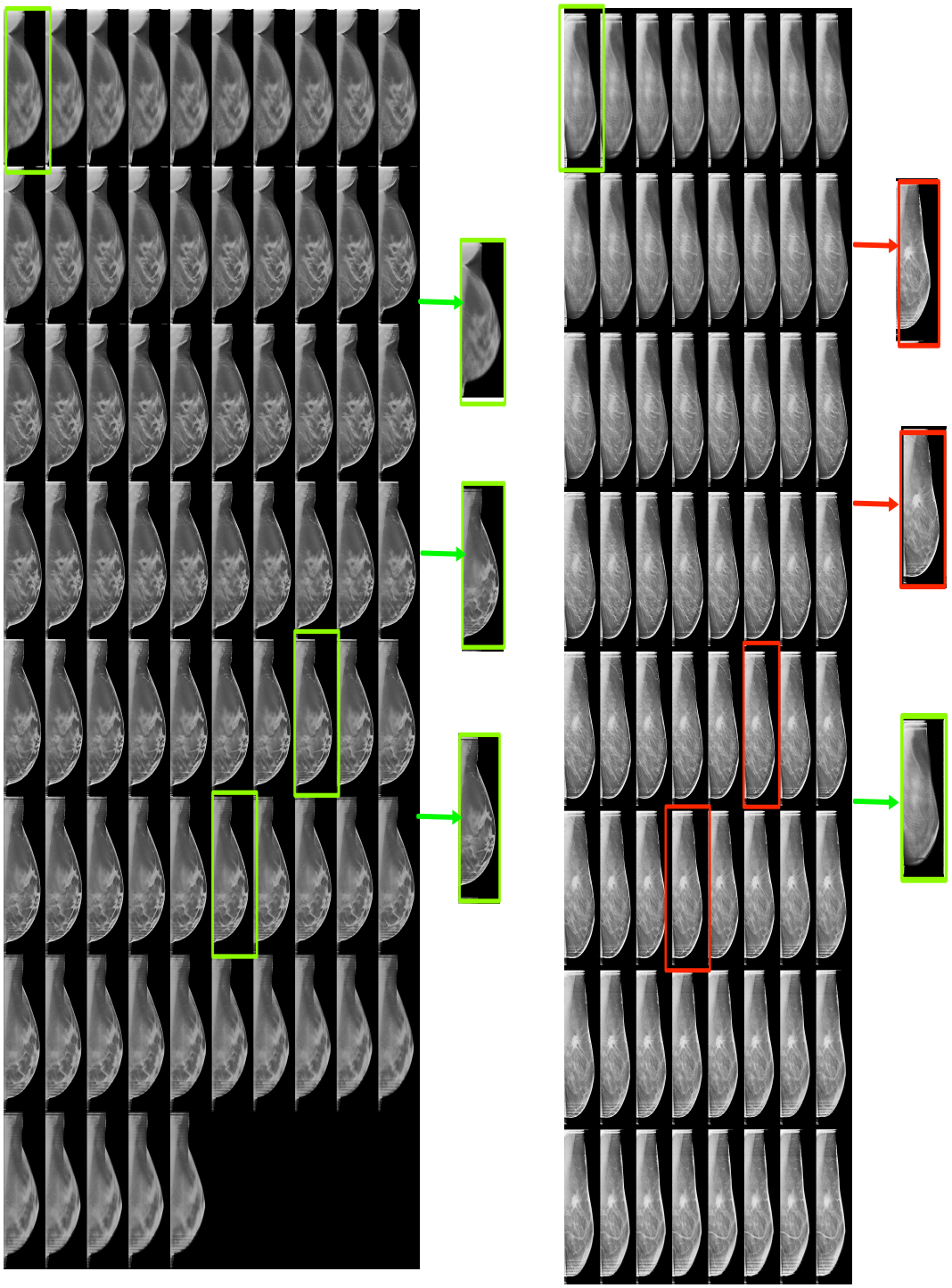


FIGURE 4.13: Sample examples for DBT cancer and non-cancer images: left image is the negative bag (non-cancer stack of 2D slices); right image is the positive bag (cancer stack of 2D slices). Red rectangles show the positive instances (cancerous 2D slices), and the green ones are negative instances (non-cancerous 2D slice).

Random Forests

The RF classification model was introduced in [5]; this approach is an ensemble of M independent decision trees f_m:

$$f_m(x) : \mathrm{X} \longrightarrow \mathrm{Y} \tag{4.32}$$

where $Y = \{1, ..., K\}$ is the set of classes. The predictive confidence for an RF classifier is defined:

$$F_k(x) = \frac{1}{M} \sum_{m=1}^{M} p_m(k|x) \quad k \in \mathrm{Y} \tag{4.33}$$

where $p_m(k|x)$ is the estimated density of kth class of the leaf from the mth decision tree. Taking the maximum overall probabilities of classes yields the output class:

$$C(x) : \mathrm{X} \longrightarrow \mathrm{Y}, \quad C(x) = \arg\max_{k \in \mathrm{Y}} F_k(x) \tag{4.34}$$

The upper bound of the error for an RF classier is defined [19]:

$$\mathrm{PE} \leq \frac{\bar{\rho}(1 - s^2)}{s^2} \tag{4.35}$$

where $\bar{\rho}$ and s are the mean correlation between pairs of decision trees and the strength of ensemble, respectively. Breiman [19] defined the margin function of an RF:

$$\mathrm{margin(x, k)} = \frac{\sum_{m=1}^{M} \mathbb{1}(f_m(\mathrm{x}) = k)}{M} - \max_{j \neq k} \left[\frac{\sum_{m=1}^{M} \mathbb{1}(f_m(\mathrm{x}) = j)}{M} \right] \tag{4.36}$$

where $\mathbb{1}$ is the indicator function and $j \in \mathrm{Y}$. The expectation of the margin in e.q 4.36 over data is called the strength term s. We can achieve the RF with the low error by reducing the correlation between trees considering that its variance is not increased too much. For reduction the correlation values among trees, each tree is provided by sub-sampling with replacement from the training data. So that for creating each tree an individual subset of the training set is used. The training data for each tress is separated into samples to be classified, in which the purity of

groups is maximized when a new node is created. The entropy measures the impurity of trees:

$$H(I) = -\sum_{i=1}^{K} p_i^j \log_2(p_i^j) \tag{4.37}$$

where p_i^j is the density of ith class in jth node. The training process of trees is continued until a maximum depth of the tree is achieved, or further information gain is impossible.

MIL-RF Settings on DBT images

As authors in [55] mentioned the obvious way to apply the MIL constraints to the RFs is adopting the MIL for each decision tree. Note that the architecture of MIL with RFs must be capable of keeping the diversity between trees to avoid incrementing the error of RFs. Hence finding the optimization strategy that preserves the proper regularization of trees in RFs plays the critical role in the efficient MI-RF architecture.

In our study, we are given the training set of bags DBT_i $i = 1, ..., n$. Each bag contains a set of learned features of corresponding 2D slice images. Therefore, $DBT_i = \left\{S_i^j, j = 1, ..., n_i\right\}$ where S_i^j is representing jth instance (2D slice) features from ith DBT (bag). The MI-RF is formulated as the optimization procedure in [55], in which the RFs is trained and labels of instances become optimization variables.

$$((y_i^j)^*, \mathrm{F}^*) = \mathrm{argmin}_{y_i^j, F(.)} \Sigma_{i=1}^{N} \Sigma_{j=1}^{n_i} \mathrm{L}(\mathrm{F}_{y_i^j}(S_i^j))) \tag{4.38}$$

$$\text{s.t.} \quad \forall i : \Sigma_{j=1}^{n_i} \mathbb{1}(y_i = \mathrm{argmax}_{k \in \mathrm{Y}} \mathrm{F}_k(S_i^j)) \geq 1$$

where $y_i^j \in$ Y and L, $\mathbb{1}$ are the loss function and the indicator function, respectively. N, n_i are the number of DBT images and amount of corresponding instances in ith bag. In this optimization procedure the RFs is trained, and simultaneously the suitable set of labels for instances are found

by minimizing the loss function L all over the entire instance samples. Thus the optimization procedure in 4.38 is the non-convex problem.

We follow the model that is proposed in [55] to solve this non-convex optimization procedure using the deterministic annealing (DA). We add a convex entropy term to the hard non-convex optimization problem. Then the optimization problem gradually reverses back to the original form using the cooling parameter T:

$$p^* = \arg\min_{p \in P} E_p(L(y)) - \mathrm{T}H(p) \tag{4.39}$$

where $L(y)$, and P are the objective function, and the space of probability distribution, respectively. The cooling parameter sequence is represented as the $T_0 > T_1 > ... > T_\infty = 0$.

DA-based Optimization MI-RF

The objective function in 4.38 is reformulated to the DA optimization format:

$$((\hat{P}, \mathrm{F}) = \mathrm{argmin}_{P,F(.)} \Sigma_{i=1}^{N} \Sigma_{j=1}^{n_i} \Sigma_{k=1}^{K} \hat{P}(k|S_i^j) \mathrm{l}(\mathrm{F}_k(S_i^j))) + T\Sigma_1^N H(\hat{P}_i) \tag{4.40}$$

$$H(\hat{P}_i) = -\sum_{j=1}^{n_i} \sum_{k=1}^{K} \hat{P}(k|S_i^j) \log(\hat{P}(k|S_i^j)) \tag{4.41}$$

The optimization procedure in e.q 4.40 consists of the following steps. First, the RFs without any MIL constraint is trained on labeled instances using their corresponding bags labels for annotation instances. Then, instance labels in target bags are considered as random variables defined over the space of probability distributions P. Afterward, the probabilities stored in the leaf of trees $\hat{P} \in P$ that are used by each tree to infer instance labels are adjusted over the bags based on the optimization problem in e.q 4.40.

Note that for the given temperature value the optimization problem is split into the two steps. First, the objective function is optimized by fixing the $\hat{P}$ distribution. Then, the distribution P^* using the entropy is adjusted over the entire bags.

Summary of MI-RF Learning on DBT

Learned features of 2D slices from the extraction paradigm (hand-crafted feature based, DCaRBM or DCNN) and corresponding DBT images are considered as the training instances and bags, respectively. First, the RFs is initialized by training it on instances that inherited their labels from DBTs where they lie. Then at a given iteration t, to train the new RF, for each tree individually a pseudo label for each instance is drawn based on the probabilities that stored in the leaf of trees from the previous iteration. If we have a positive bag (DBT) with no instance (2D slice) labeled as positive, the positive pseudo-label is assigned to the instance with the highest probability of being positive. Trees are trained on pseudo labeled instances individually. This training procedure is continued until the temperature value reaches the minimum value. The cooling parameter T is scheduled with function $\mathrm{T}(t) = e^{-t.C}$ where t expresses the current iteration and $C = \{0.1\}$ is a constant.

4.4.6 Other MIL Models

In this study, we like to examine another popular MIL approaches in the CAD frameworks based on the synergy of support vector machines (SVM) and MIL. The SVM approach is one of the most common supervised learning methods for different learning tasks. The standard SVM is a non-probabilistic binary classifier that provides the separation hyperplane between the examples from the two classes, in which this hyperplane separates samples with the highest possible margin.

The binary MIL models our problem $DBT_i = \left\{S_i^j, j = 1, ..., n_i\right\}$ with each bag DBT_i a label y_i is associated. These labels are interpreted as: if $y_i = -1$, all $y_i^j = -1$. The relation between instance labels and y_i^j and bag labels y_i is defined as $y_i = \max_{j \in n_i} y_i^j$ while it can be presented in term of linear formulation:

$$\sum_{j \in n_i} \frac{y_i^j + 1}{2} \geq 1, \;\; \forall i \,\text{s.t.}\, y_i = 1, \;\; \text{and}\, y_i^j = -1, \, \forall i \,\text{s.t.}\, y_i = -1 \tag{4.42}$$

The mixed integer SVM (mi-SVM) is formulated using the SVM constraints while this method leaves the label of instances unknown in positive bags. It applies the constraint for positive bags that guarantees at least one positive instance is available in the bag. The formulation of generalized soft-margin mi-SVM is given by:

$$\begin{aligned} &\min_{y_i^j,\mathrm{w},b,\xi} \frac{1}{2}\mathrm{w}^T\mathrm{w} + C\sum_{i\in N}\sum_{j\in n_i}\xi_{ij} \\ &\text{s.t. } y_i^j(\mathrm{w}^T\phi(S_i^j)+b) \geq 1-\xi_{ij}, \\ &\xi_{ij}\geq 0, \quad y_i^j \quad \in \{-1,1\}, \\ &\begin{cases} \sum_{j\in n_i}\frac{y_i^j+1}{2} \geq 1 & \text{if} \quad y_i = 1 \\ y_i^j = -1 & \text{if} \quad y_i = -1 \end{cases} \end{aligned} \tag{4.43}$$

where the function $\phi(S_i^j)$ maps the instances S_i^j to the higher dimensional kernel space, $C > 0$ is the regularization parameter, and $0 < \xi_{ij} \leq 1$ are margin violation points to allow classification errors. The mi-SVM is a mixed problem that should find the optimal labels for instances and the optimal hyperplane. In mi-SVM, the soft-margin criterion is maximized over all possible labels as well as hyperplanes.

Another way to applying the maximum margins to the MIL setting is to employ the margin to bags instead of instances concerning the hyperplane:

$$y_i \max_{i\in N}(\mathrm{w}^T\phi(S_i^j)+b) \tag{4.44}$$

This approach selects the instance with the most positive instance of the positive bags and the instance with the least negative labels of the negative bags for representing bags. E.q 4.43 is rewritten with the functional bag margin in 4.44 and gives the MI-SVM formulation as:

$$\begin{aligned} &\min_{y_i,\mathrm{w},b,\xi} \frac{1}{2}\mathrm{w}^T\mathrm{w} + C\sum_N \sum_{i\in N} \xi_i \\ \text{s.t.} \quad & y_i \max_{i\in N}(\mathrm{w}^T\phi(S_i^j) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \quad y_i \in \{-1, 1\} \end{aligned} \tag{4.45}$$

Another SVM-based MI approach is the multiple-instance classification algorithm (MICA). In the MI-SVM algorithm, a particular instance is chosen in each bag that acts as the witness. But in the MICA approach, the witness is found using the linear programming in which it finds the arbitrary combinations of points in a positive bag that serve as the witness [32].

Multiple instance semi-supervised SVM (MissSVM) [32] is like mi-SVM by semi-supervised SVM approaches. The optimization procedure is the same as the MI-SVM optimization with two more constraints that enforce a positive instance and a negative one in the positive bags.

Single-instance learning (SIL) is the most straightforward approach to estimate the instance-level classifier. It assigns the label of the bags to their instances. For this classifier, the supervised learning classifier (SVM in this study) with mislabeling negative instances in the positive bags is used to model the MIL learner. The SIL formulation is given in:

$$\begin{aligned} &\min_{y_i,\mathrm{w},b,\xi} \frac{1}{2}\mathrm{w}^T\mathrm{w} + C\sum_{i\in N} \sum_{j\in n_i} \xi_{ij} \\ \text{s.t.} \quad & y_i(\mathrm{w}^T\phi(S_i^j) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \quad y_i \in \{-1, 1\} \end{aligned} \tag{4.46}$$

Optimization Procedure

Likewise the optimization procedure in an SVM method, the optimization problem is the dual form that maximizes the margin hyperplane. First the solution for w, and b are found:

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\alpha^T\mathrm{Q}\alpha - \mathrm{e}^T\alpha \\ \text{s.t.} \quad & \mathrm{y}^T\alpha = 0 \quad \text{and} \quad 0 \leq \alpha \leq C \end{aligned} \tag{4.47}$$

where e, and $\mathrm{Q}_{mn} = y_m y_n K(x_m, x_n)$ are the vector of ones, and the positive semi-definite matrix, respectively, and $K(x_m, x_n)$ is the kernel function. By solving this dual problem, w and the bias value b of the hyperplane are found. Now the hyperplane is available, then using this hyperplane the instances are classified to obtain new labels. We apply the MIL constraint that at least one positive label is available in each bag. By chance, if we find some positive bags that do not include the positive instance, we assign positive pseudo-label to the instance that is the closest to the hyperplane. The optimization procedure is continued until the algorithm converges to labels so that labels of the instances do not change anymore.

Chapter 5

Experimental Evaluations

5.1 Datasets: Synthetic DBT Dataset, Real DBT Datasets

In this thesis, we validate and compare our CAD frameworks on the DBT medical image the dataset which includes the real data, and the synthetic data. In medical imaging studies, it is common to use the simulated data that provided from the real data. The reason is that the real datasets with confirmed lesions especially for DBT images are very limited. In this section first, we introduce the real dataset. Then we discuss our technique for generating more DBT data.

5.1.1 DBT Dataset

The benchmarking clinical DBT images used were provided by medical doctors in the Breast Imaging Research Laboratory of at Massachusetts General Hospital (MGH). Professor Richard Moore was our project contact at MGH; he developed algorithms for breast imaging technology. He is a partner with Dr. Kopans, a patent holder of the DBT. Each DBT was generated by low-dose exposures, in which the total shot doses should be in the range of regular mammogram dose. DBT cases were acquired in a mediolateral oblique (MLO) view by a prototype General Electrics. The system used an X-ray source that covered an arc of 50^o and provided eleven projection views. DBTs were reconstructed to 1-mm slices spacing and saved in SMV format. The images are large volume images (300-500 MB). We converted SMV images into TIFF stack/slices and used data in TIFF format. The number of provided 2D slices varied based

on breast thickness between 50 and 80 slices per breast. Each 2D slice with the resolution of 1270×2304 pixels was saved in TIFF format. The dataset contains 87 DBT images from 87 individual breasts, of which 27 are malignant and 60 were normal. Non-cancerous DBT volumes without early signs of breast cancer were obtained from 30 patients, in which images were taken from the left and right breasts. The cohort of cancerous cases was annotated by two Mammography Quality Standards Act (MQSA) radiologists with a 2D bounding box for DBT (see Figure 5.2). The dataset includes invasive ductal carcinoma (IDC), the most common type of breast cancer; ductal carcinoma in situ (a common, non-invasive breast cancer); invasive carcinoma; and invasive lobular carcinoma cases (see Figure 5.3). The dataset includes 87 DBT volumes with corresponding $5{,}040$ 2D X-ray images. Figure 5.1 illustrates an example of 2D slices that are extracted from the raw 3D reconstructed DBT image.

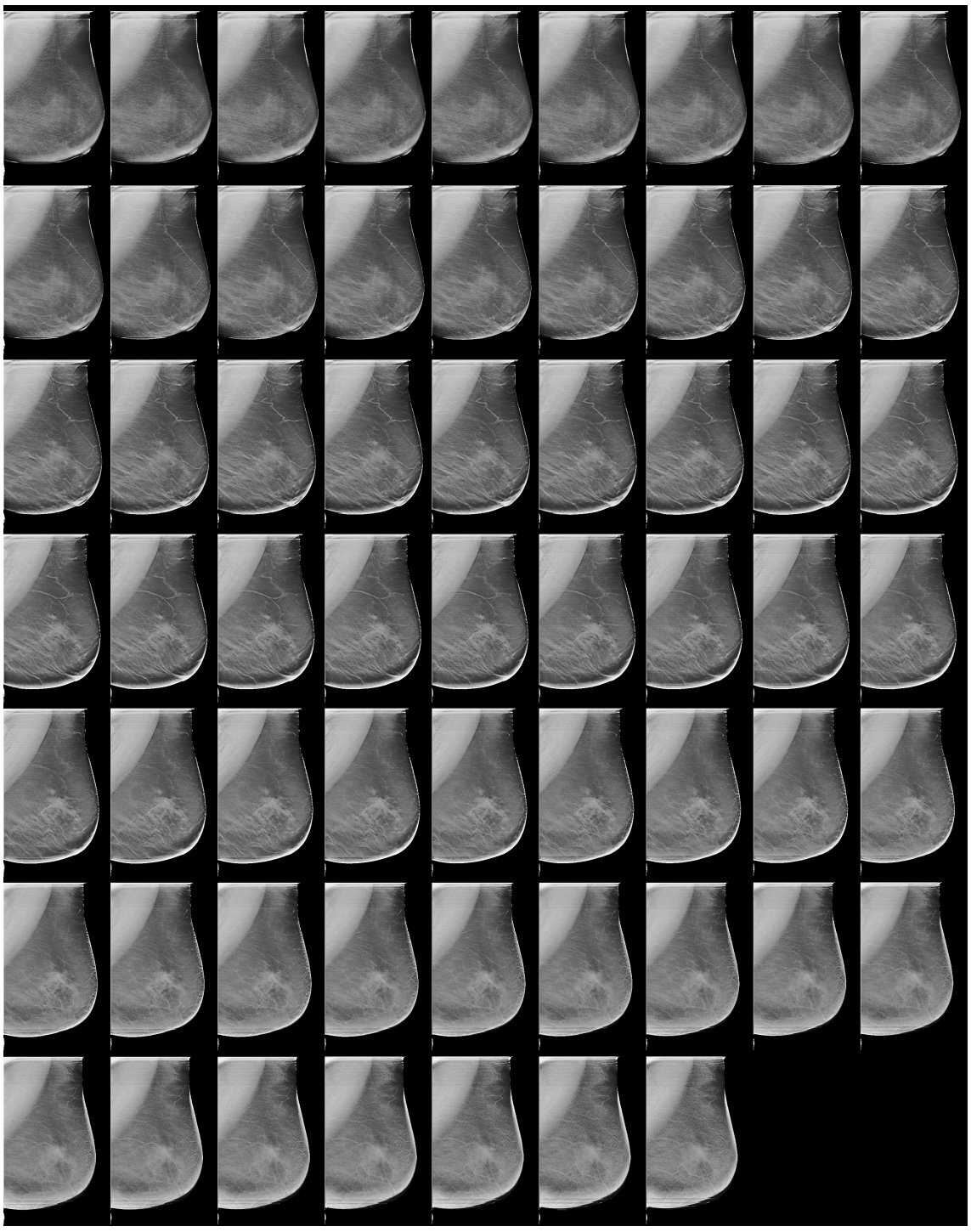


FIGURE 5.1: 61 2D image slices that are extracted from the quasi 3D DBT volume.

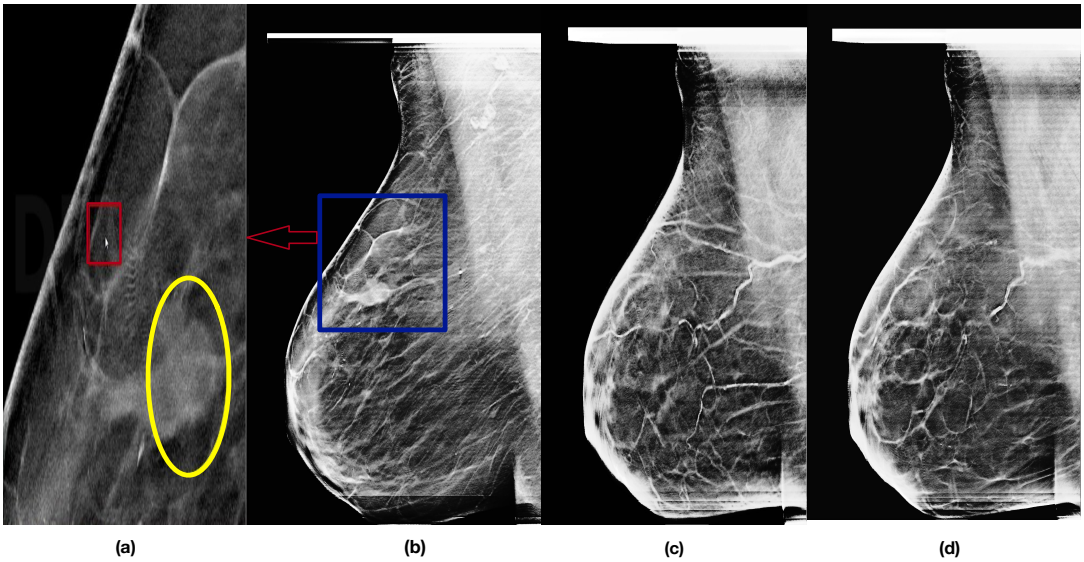


FIGURE 5.2: (a) Irregular spiculated masses highlighted in the red box and the yellow circle on the slice 15 (b) of a cancer DBT case, whereas they do not appear in (c) and (d) slices of the DBT.

5.1.2 Data Augmentation

As mentioned earlier, we developed three CAD frameworks for detecting breast cancer in DBT. Two of them utilize deep learning models, DCNN and DCaRBM architectures. These models are powerful neural network models that can memorize the training data. Deep learning architectures have many hidden layers with a vast number of parameters. If a model has enough parameters, the model performance on the training data may approach perfection, while its performance on the testing data decreases. This issue is called over-fitting and is the significant drawback of the training deep learning models. The over-fitting problem causes the model reduces the ability of generalization. In the other word, there is not enough data for building the model. For example, the authors in [76] proposed a deep learning CAD framework on DBT images based on the transfer learning technique using the large mammography data. This model was trained on the extracted ROIs of mammography data. Then all weights in the first three convolutional layers were frozen, and the last convolutional layer and fully- connected layers were initialized and trained with DBT training ROIs. In another study, synthetic X-ray projections were generated by simulation models [41]. The proposed simulation model used algorithmic and software approaches to simulate 3D breast images with micro-calcifications. Though various studies have

PATH	1S	1X	1Y	1Z	S2	2X	2Y	2Z	PATH
M	R	328	1286	17					INVASIVE DUCTAL CARCINOMA
M	L	886	932	34					INVASIVE DUCTAL CARCINOMA
M	L	1202	1397	20					INVASIVE DUCTAL CARCINOMA
M	L	458	1108	27					DUCTAL CARCINOMA IN SITU
M	L	560	1263	51					INVASIVE CARCINOMA
M	L	615	1048	15					INVASIVE LOBULAR CARCINOMA
M	R	339	1113	52					INVASIVE CARCINOMA
M	R	723	1474	62					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	R	747	1187	33					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	R	601	1193	13					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	L	918	1573	9					DUCTAL CARCINOMA IN SITU
M	L	500	1158	14					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	L	1173	1540	61					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	R	123	1441	16					DUCTAL CARCINOMA IN SITU
M	L	355	965	25					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	L	810	1474	30					DUCTAL CARCINOMA IN SITU
M	R	489	1284	45					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	L	597	1676	11	L	259	1698	11	INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	R	328	871	45					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	L	1059	1040	19					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	R	577	1243	33					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	L	396	1306	20					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	L	466	1203	23					INVASIVE LOBULAR CARCINOMA, DUCTAL CARCINOMA IN SITU
M	L	538	1138	35					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU
M	R	548	1178	26					INVASIVE DUCTAL CARCINOMA, DUCTAL CARCINOMA IN SITU

FIGURE 5.3: DBT dataset annotated by radiologists. M in the left PATH column represents the malignant lesion. The right PATH column shows the labels for different types of breast cancers. The S columns include R and L to show the lesion is on the right and left breast, respectively. X, Y, and Z are the lesion coordinates on the DBT image.

investigated and authors have introduced many methods in the literature, reduction the over-fitting is still an open subject of research in the machine learning area.

In this study, we employed various techniques to avoid the over-fitting issue for our deep learning models. We applied some methods to design of deep learning models that mentioned in 4. Such as the dropout layer in DCNNs, and the application of the sparsity to design CaRBMs. Moreover, we cropped out the small randomly parts from images and treat them as the input of networks for training deep learning models[51]. Furthermore, in this section, we introduce the data augmentation techniques to generate artificially more 2D X-ray images to train deep learning CAD models.

Data Augmentation Approaches on the DBT Data

Geometrical transformations are categorized into the rigid, and the non-rigid transformations (see Figure 5.4). The rigid ones including the translations and rotations keep all distances constant. In non-rigid transformations such as polynomial, thin-plate spline interpolation, and B-spline, distances changed, but the lines are kept straight. In this thesis, we utilize both rigid and non-rigid transformations to enlarge our data for training the deep learning based CAD frameworks.

2D slices of DBT images are rotated in $[0^0, 90^0, 180^0, 270^0]$ [47]. This provides 4 DBT slices from each 2D slice. Then the non-rigid thin-plate spline transformation is applied on 2D slices.

After applying the rotating and flipping techniques, we use thin-plate spline (TPS) deformation on the DBT slices. First, we need to extract fiducial points of 2D slices. First, we obtain the contour of the breast in each 2D slice by a simple thresholding method, then 2D fiducial points ($\{F_i, i = 1, .., k\}$) are selected through a sampling step along the breast contours and left the border of the 2D slice images. Extracted fiducial points are transformed randomly at the 2D slice, and a deformed image is generated using:

$$f(x, y) = a_1 + a_x x + a_y y + \sum_{1}^{k} w_i U(|F_i - (x, y)|) \tag{5.1}$$

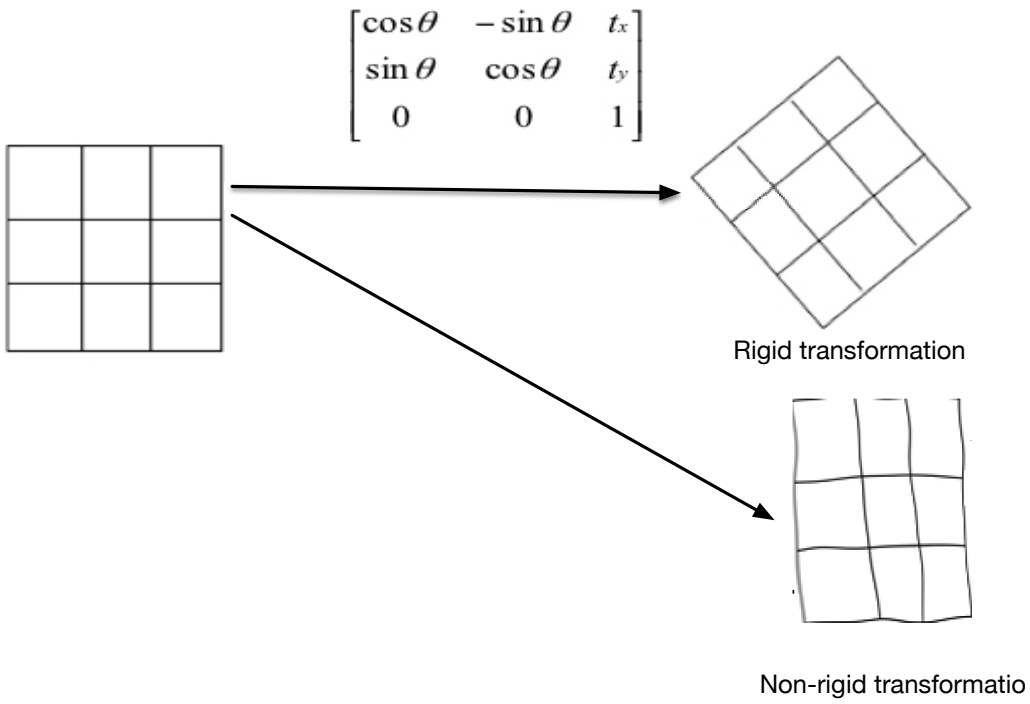


FIGURE 5.4: The input image (left image) the rigid, and non-rigid transformed images (right images).

where $f(x, y)$ denotes the displacement of an arbitrary 2D point (x, y), and a_1, a_x, a_y, and w_i represent spline coefficients. The final summation term controls the bending influenced by k fiducial points using a controlling function $U(r) = r^2 \log(r)$, where r is the Euclidean distance $\sqrt{((x_i - x)^2 + (y_i - y)^2}$ from the ith fiducial point to (x, y). In fact, TPS defines a surface on fiducial points, while the function $f(x, y)$ minimizes the bending energy with $\triangle^2 f = 0$.

After application data augmentation techniques to our dataset, the DBT data was enlarged to 696 DBT volumes with 40,320 corresponding 2D slices. Note that we must increase the size of the Data for training deep-learning based CAD frameworks. For the hand-crafted, feature-based CAD framework, we use the original data to extract features and learn image patterns.

5.2 Performance Metrics

We discuss the evaluation concepts used for the performance analysis of three CAD frameworks for detection spiculated masses in DBT images in the following subsections.

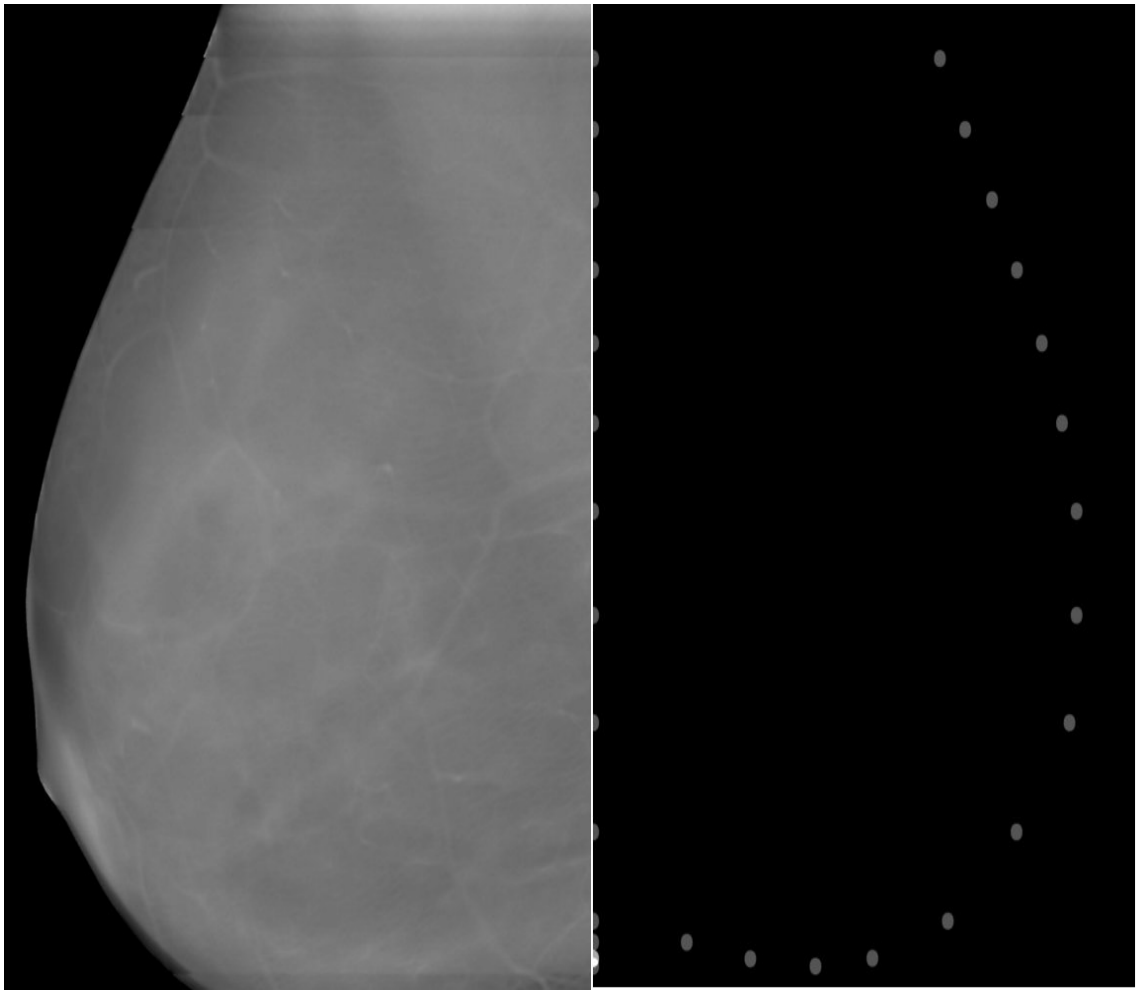


FIGURE 5.5: (a) A 2D slice sample, and (b) selected fiducial points from (a).

5.2.1 Sensitivity and Specificity

A classification model maps the instances into the available labels for the input data such as diseased or healthy classes for medical images. The classifiers on medical images are usually binary classification so that they result in a positive (p) or a negative (n) decision. Considering a set of ROIs or the whole images, some of them exhibiting normal tissue (n) and others with signs of abnormality (p). The correct type of each image is referred to as the ground truth.

There are four possible outcomes for binary classifications. If the result forms a prediction p and the correct label is p, this is called true positive (TP). If the prediction value is p while its label is n that is called the false positive (FP). Contrariwise true negative (TN) happens if predicted and

actual labels are negative; if the predicted class is positive, it is called false positive (FP).

Consider a CAD framework developed to classify medical images as positive, and negative cases. To evaluate the performance of a CAD framework common metrics are sensitivity and specificity. The sensitivity is defined as the fraction of positive samples that are annotated as positive, and the specificity is the fraction of negative ones that are identified as negatives by the framework.

Sensitivity indicates how sensitive is the system for detection of positive samples. The high sensitivity rate shows that the framework very rarely misses an abnormal (p) samples. The most reliable CAD framework is that one the has the highest sensitivity. The high specificity means the system rarely classifies the normal (n) sample as the abnormal (p) cases.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TN+TP+FP+FN}}$$

False-positive ratio (FPR) represents the rate of the false abnormal detection samples. It is essential for a CAD framework to produce as low as possible FPR. This problem that is called false alarm may increase the anxiety and health-care costs to the patient.

$$\text{FPR} = 1 - \text{specificity}$$

False-negative ratio (FNR) represents the rate of the missed abnormal samples. This issue delays the treatment for patients and in the worst case scenario may lead to the patient's death.

$$\text{FNR} = 1 - \text{sensitivity}$$

5.2.2 Receiver Operating Characteristic (ROC) Analysis

A ROC curve is another evaluation method. This curve is a graph of the true positive rate versus false positive rate are obtained by varying a threshold for a binary classifier system. It is plotted in the unit square, where the true positive rate on the ordinate (y-axis), and the false positive rate on the abscissa (x-axis). The ideal point for the CAD framework sensitivity and specificity are one; this point is at $(0, 1)$. ROC analysis can provide tools to select the optimal models. The area under the ROC curve (AUC) is the measurement of the overall performance of a system. The desired CAD framework results in AUC=1.

5.3 Mass Detection Results

We validated and compared three CAD frameworks (the DCNN-, DCaRBM-, and hand-crafted feature-based frameworks using the MI-RF classifier) on the DBT medical image dataset. Furthermore, we evaluated the MI-RF learner through four other MIL-based classifiers: the MISVM, SIL, MissSVM, and MICA. Consequently, we examined a total of fifteen frameworks for DBT image classification. The results are reported three in Table 5.5 (CAD with MI-RF classifier), three in Table 5.6 (CAD with MI-SVM classifier), three in Table 5.7 (CAD with MissSVM classifier), three in Table 5.8 (CAD with SIL classifier), and three in Table 5.9 (CAD with MICA classifier). The best results of these frameworks were published in [104].

We developed three pipelines for three CAD frameworks. The whole implementation of three CAD frameworks is in Python and C++ languages. The main approaches in the hand-crafted feature-based CAD framework for detecting spiculated masses are Gabor wavelet transformation, cooperative of a principled deep random field model and graph cut segmentation and MI-RF learning. The Skimage Python library, the QPBO C++ scripts, and also graph cut algorithms

such as Min-Cut/ Max-Flow algorithms for energy minimization C++ scripts have been used to implement mass area detection and segmentation approaches. The DCaRBM MI-RF CAD framework consists of DCaRBMs performed by the C++ library and also the proper adjustment that we applied to the model for mass detection on DBT images. For implementing DCNN, we used the Tensorflow library with GPU support specifically the NVIDIA GTX 1080 11 GB GPU with CUDA toolkit. The MI-RF classifier was implemented in the C++ library, we modified this library based on our data and wrote the Python wrapper for the whole library. In all these experiments we use these frameworks for the mass detection problem in DBT images. However, as mentioned earlier, we also substitute the MI-RF with other MIL classifications designed utilizing the SVM classier. The CVXOPT library was used to solve the quadratic programs (QPs) in SVM based models. For both MI-RF and multiple instance SVM based models, we wrote the scripts that fetch information from the list of 2D slices of DBT images.

The DBT images were obtained from MGH hospital. The dataset consists of the 87 views from 87 breasts, of which 27 were malignant and 60 were normal. Non-cancerous DBT volumes without early signs of breast cancer were obtained from 30 patients. These images are large volume (300$\sim$500 MB) and including (50$\sim$70 slices). The data set includes 87 DBT volumes with 5,040 corresponding 2D X-ray images. The radiologist professor Richard Moore was our project contact at MGH; he annotated the images. Due to DCNN- and DCaRBM- based frameworks require a large amount of labeled data for training, we artificially generated more 2D X-ray images to train the DCNN and DCaRM models. In the two-step enlargement technique, first, each DBT 2D slice was rotated; then, the thin-plate spline (TPS) deformation was applied to the DBT slices. Following data augmentation, the number of 2D slices used for training increased from 5,040 to 40,320. The details of the data set are reported in Table 5.1.

We utilize the MI-RF in the CAD architectures as the main learner procedure to deliver multi-class instance classifier. Using this learner in CAD frameworks make them capable of detecting other early sign of breast cancer on DBT images with new clinical cases. In this experiment, we employ it for two-class prediction problem (spiculated mass, and normal classes). Note that the SVM based multiple instance baselines work for binary classification, but we use these algorithms for multi-class problems (mass- Normal). We evaluate the SVM baselines in CAD

TABLE 5.1: Data used to extract information from 2D slices and 2D and DBT data MIL training. Note: For the training set of deep learning based models on 2D slices, each slice was augmented, resulting in eight slices. For the MIL testing and training sets, we did not use augmented data.

CAD framework	No. of 2D data for learning features	No. of 2D instances in MIL	No. of bag data in MIL
DCNN based	40320	5040	87
DCaRBM based	40320	5040	87
Hand-crafted feature based	5040	5040	87

systems with training a set of masses versus the normal class cases. To evaluate each framework, performance metrics based on TP, TN, FP, and FN are calculated separately (see Table 5.2).

TABLE 5.2: Performance metrics for mass detection in the DBT datasets.

Measure	Description
TP	CAD framework=mass, radiologist=mass
FP	CAD framework=mass, radiologist=benign
TN	CAD framework=benign, radiologist=benign
FN	CAD framework=benign, radiologist=mass
Sensitivity	TP/(TP+FN)
Specificity	TN/(TN+FP)
Accuracy	(TP+TN)/(TN+TP+FP+FN)
AUC	Area under receiver operating characteristic (ROC) curve

The preprocessed input data was passed to the feature paradigm of each CAD frameworks for obtaining information from 2D slices. The deep learning-based systems were trained on the input data to capture features from samples. For training deep learning models we used 70% of the data for the training, and 30% for testing sets. Note that we maintained the original image and its corresponding augmented images in the same set (training or testing). In the learning task,

the features extracted from 2D slices were conveyed to the MIL learner for categorizing DBT volumes into two classes: spiculated mass and normal cases. The details of the data used to extract information from 2D slices as well as the training data for MIL classifiers are reported in Table 5.1. Bag data including learned features from 2D slices are split into two subsets, in which the proportion of malignant and normal cases are maintained. The training set comprised of 14 masses and 15 normal cases, and the test set contained 13 mass cases and 15 normal samples. We performed five random training and testing sets partitions. The partition that obtained the best ROC result for each CAD framework is used to report results and plot performance curve.

5.3.1 MI-RF Results

We first evaluate the DBT classification performance of three CAD systems with the MI-RF classifier on DBT testing data. For five partitions, sensitivity, specificity, and global accuracy are determined, and a ROC curve is obtained, by varying a threshold on the bag testing mass likelihood scores. We use the partition that is achieved the best ROC curve to report DBT classification performance for the three CAD frameworks. The parameters for MI-RF are optimally chosen individually for each CAD framework. Table 5.3 reports various parameters for MI-RF for different CAD frameworks.

DBT classification performance for the three CAD frameworks are reported in Table 5.4. We also illustrate DBT classification performance using ROC curves (see Figure 5.6). Due to the small number of malignant case, note that the ROC curves in Figure 5.6 would not be generalized to the population.

TABLE 5.3: MI-RF parameters are used in CAD frameworks.

MI-RF parameters	DCNN based	DCaRBM based	Feature based
No. of trees	100	100	100
Train-sampling method	Bagging without refine	Bagging without refine	Bagging with refine
Max depth of trees	20	20	15
Min No. of samples split	5	5	5
Min temperature value	0.1	0.1	0.1
Max temperature value	5	5	5
Cooling-parameter-tree	Entropy	Entropy	Entropy
Loss function	Entropy	Entropy	Entropy

TABLE 5.4: Performance comparisons of the three CAD frameworks using the MI-RF classifier on DBT tests.

Measure	DCNN MI-RF	DCaRBM MI-RF	Hand-crafted Feature MI-RF
AUC	0.87	0.70	0.75
Accuracy	86.81%	78.5%	69.2%
Specificity	87.5%	66.6%	75%
Sensitivity	86.6%	81.8%	66.6%

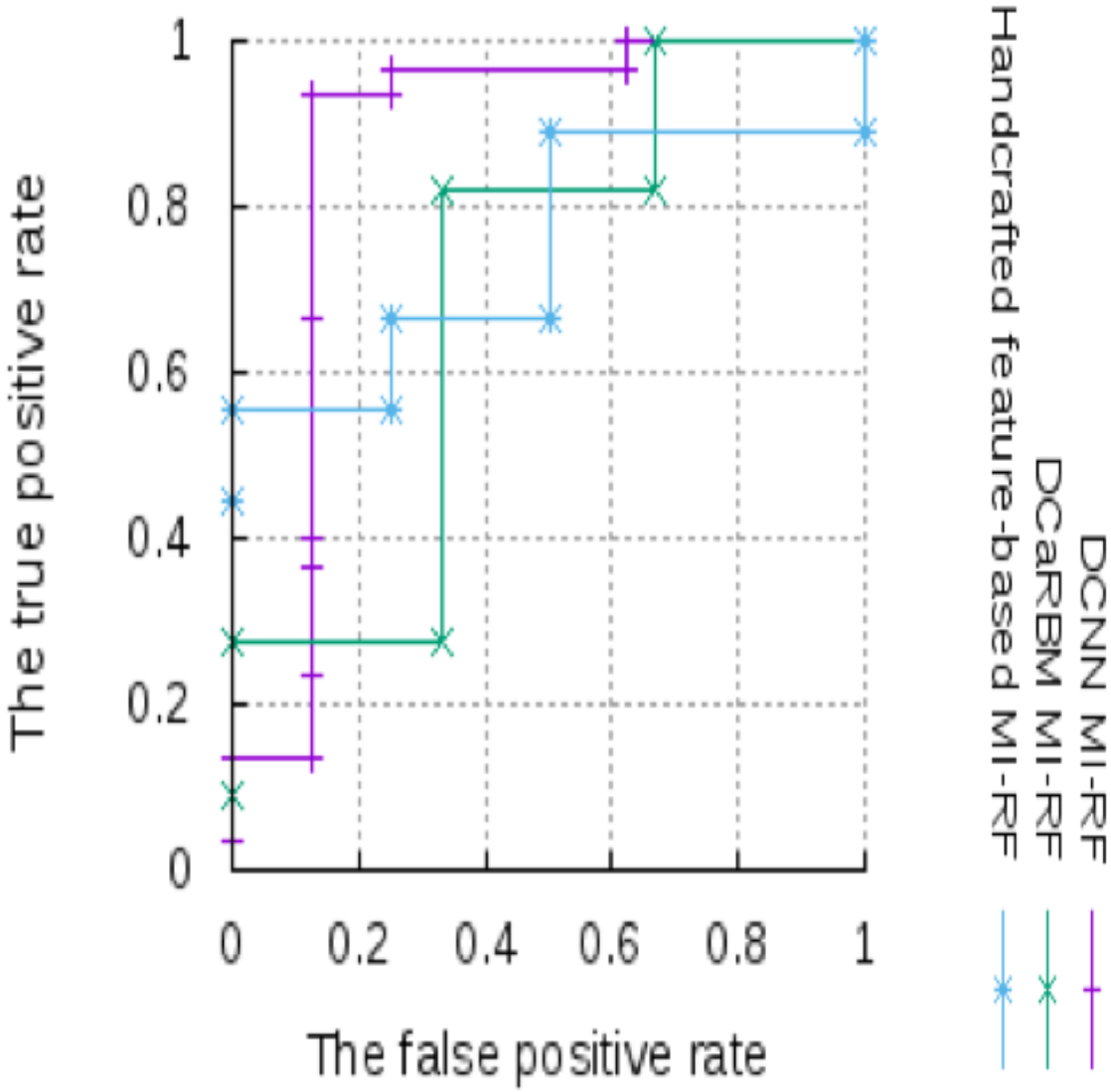


FIGURE 5.6: ROC curves for the three CAD MI-RF based frameworks DBT classification for mass detection on the DBT testing data.

5.3.2 Multiple Instance SVM Learning Baselines Results

In this section, three CAD frameworks with MICA, MI-SVM, SIL, and MissSVM classifications are evaluated for mass detection on DBT images. We use support vector machine multiple instance learners as matching templates in [32] for all experiments in this section. The parameters for these models are optimally chosen individually for each CAD framework. We use the radial basis function (RBF) kernel for all multiple-instance SVM based models. Parameters for the RBF kernel γ from $[10^{-6}, 10]$, and $C \in [10^{-2}, 10^{5}]$ are found for each CAD framework using the random parameter search method [32]. The maximum iterations for quadratic programs are set at most 100 times. For CAD frameworks based on the MICA approach, we use the L_1 regularization.

DBT classification performance of the three CAD systems with SVM based multiple-instance classifiers was evaluated using the test data. The reported results for the CAD frameworks are presented in Tables 5.5, 5.7, 5.8, and 5.6. Figures 5.7, 5.9, 5.10 and 5.8 show ROC curves for the three CAD frameworks with multiple-instance SVM classifiers.

TABLE 5.5: Performance of three CAD frameworks using the MI-SVM classifier in the DBT dataset over the tests for each system.

Measure	DCNN MISVM	DCaRBM MISVM	Hand-crafted Feature MISVM
AUC	0.72	0.65	0.52
Accuracy	69.2%	61.5%	35.71%
Specificity	66.6%	66%	33.3%
Sensitivity	70%	60%	36.36%

TABLE 5.6: Performance of the three CAD frameworks using MissSVM classifier in the DBT dataset over the tests for each system.

Measure	DCNN MissSVM	DCaRBM MissSVM	Hand-crafted Feature MissSVM
AUC	0.68	0.75	0.73
Accuracy	69%	66%	64.28%
Specificity	66%	66%	66.6%
Sensitivity	70%	67.6%	63.63%

TABLE 5.7: Performance of the three CAD frameworks using SIL classifier in the DBT dataset over the tests for each system.

Measure	DCNN SIL	DCaRBM SIL	Hand-crafted Feature SIL
AUC	0.5	0.7	0.4
Accuracy	61.5%	66.7%	35.3%
Specificity	66%	66.6%	33.3%
Sensitivity	60%	66.6%	37.5%

TABLE 5.8: Performance of the three CAD frameworks using MICA classifier in the DBT dataset over the tests for each system.

Measure	DCNN MICA	DCaRBM MICA	Hand-crafted Feature MICA
AUC	0.55	0.61	0.58
Accuracy	53.38%	53.3%	64.7%
Specificity	66%	0%	66.6%
Sensitivity	50%	66.6%	62.5%

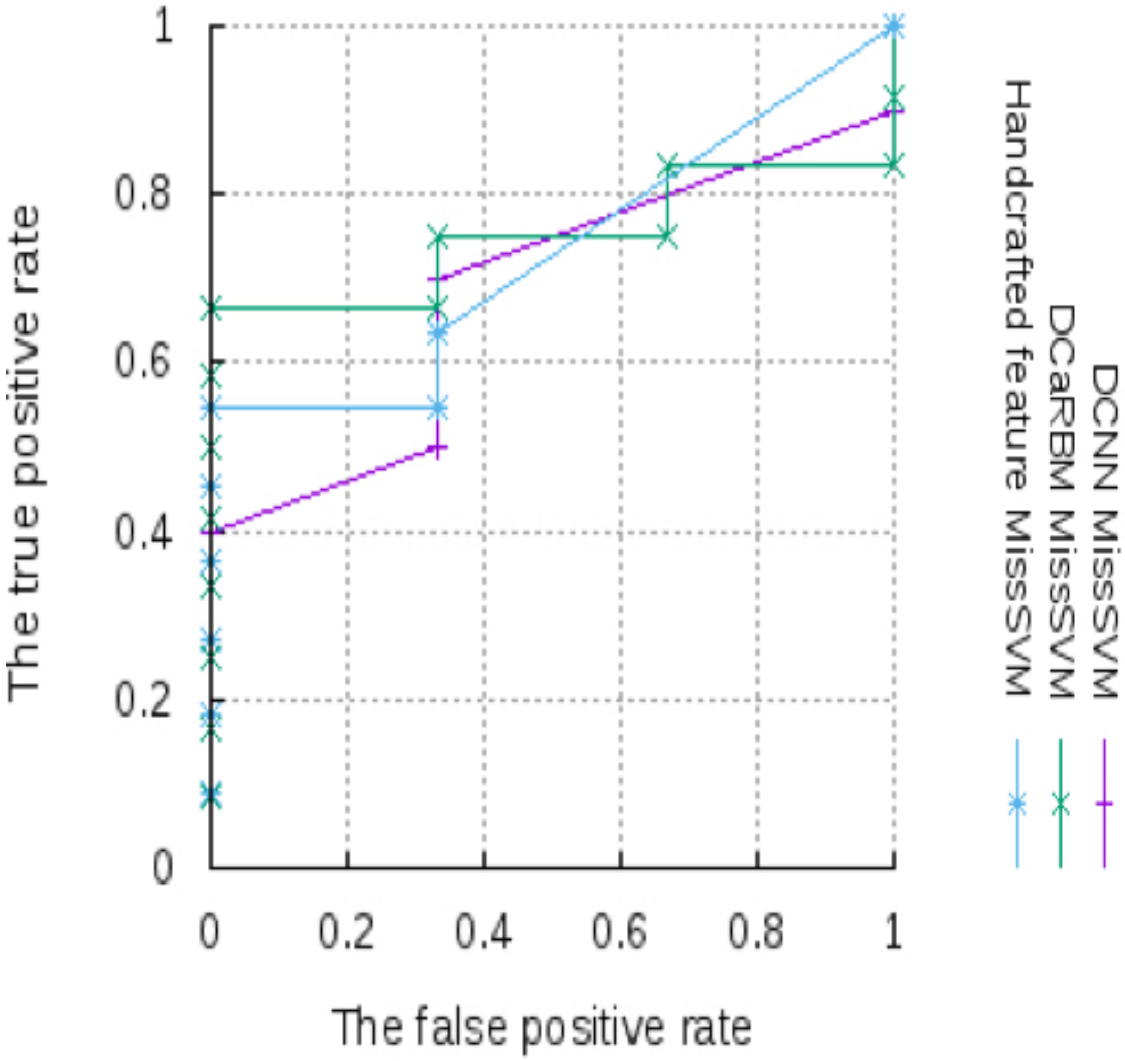


FIGURE 5.7: ROC curves for three CAD MissSVM-based models DBT classification for mass detection on the DBT test.

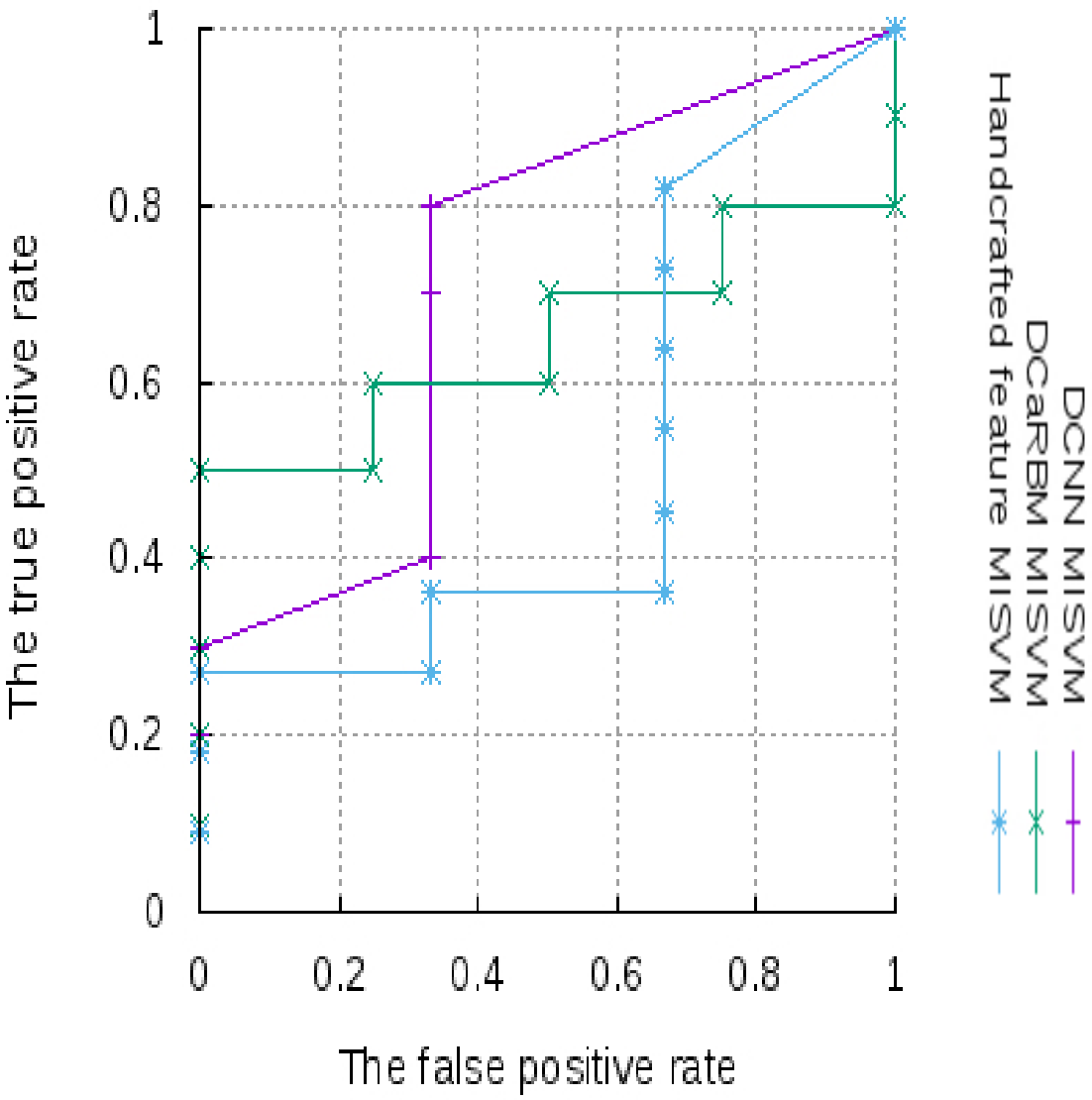


FIGURE 5.8: ROC curves for three CAD MI-SVM-based models DBT classification for mass detection on the DBT test.

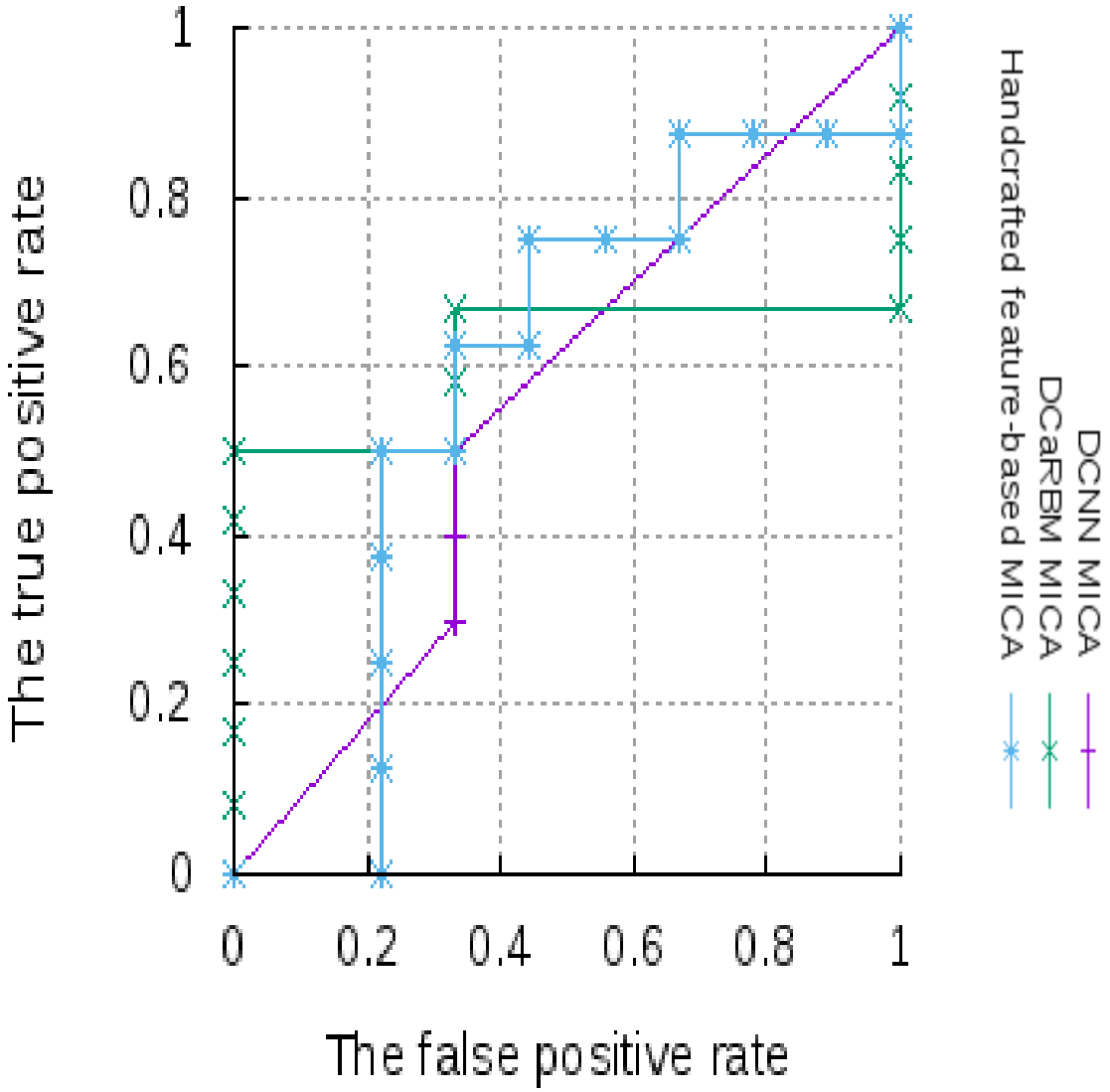


FIGURE 5.9: ROC curves for three CAD MICA-based models DBT classification for mass detection on the DBT test.

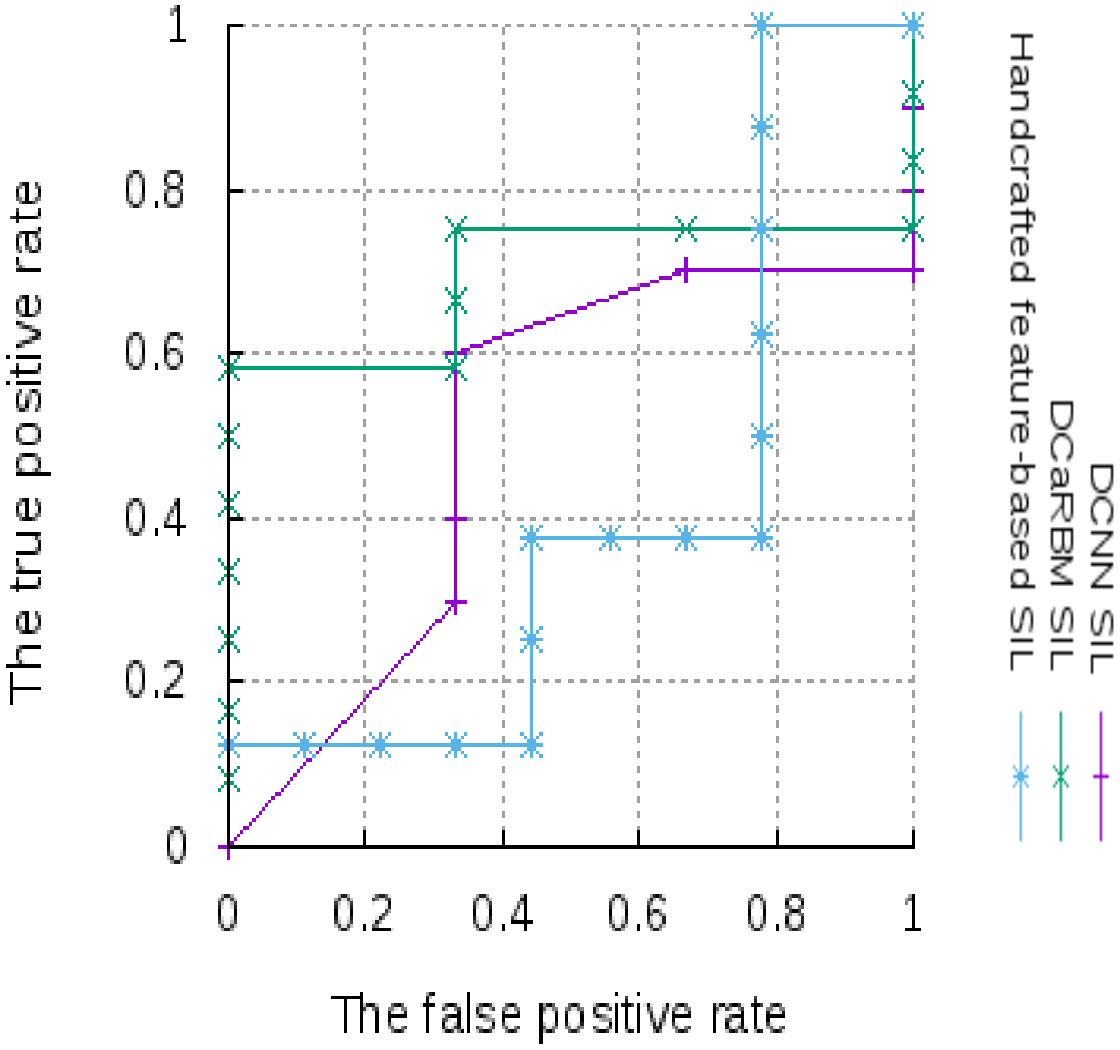


FIGURE 5.10: ROC curves for three CAD SIL-based models DBT classification for mass detection on the DBT test.

Chapter 6

Discussion and Conclusions

6.1 Discussion

We introduced three computer-aided detection frameworks to detect breast cancer on digital breast tomosynthesis images. These three CAD systems for mass detection in DBT utilized identical input and multiple-instance learning classification structures, but different feature representation paradigms for 2D slices of DBT images. For designing the first framework, we employed the hand-crafted feature extraction for detecting spiculated masses in DBT images. As opposed to the hand-crafted feature-based model, we introduced two CAD frameworks using the deep learning techniques to learn automatically hierarchical features from images. We used the deep cardinality restricted Boltzmann machines comprised the stack of three CaRBMs in the former deep model for detecting spiculated masses. For the latter deep model, we utilized the deep convolutional neural networks model with three convolutional and two fully connected layers. In the learning task, features were extracted from 2D slices, and then the MIL learner annotated DBT volumes. For the feature extraction, the hand-crafted, feature-based model required the segmentation and feature extraction stages and the deep learning models needed to be trained using the original data, and the simulated data.

In this study, our dataset in hand included 87 DBT volumes whose number of 2D slices ranges among (50-80) results in $5{,}040$ 2D images. We applied the augmentation techniques on each 2D slice by rotating and flipping methods and using the thin-plate spline algorithm, resulting in 8 2D slices to enrich the dataset for the training deep learning models. We used 70% of the

data (the original data, and the simulated ones) for the training set and 30% for the testing set. After extracting information from 2D slices either using hand-crafted features or deep learning models, for each system individually the weakly multiple-instance learning classification was trained on the information obtained to annotate DBT, where each bag in the training process comprised the learned features from 2D slices. DBT classification performance of the three CAD frameworks with five MIL methods was evaluated using the metrics of accuracy, AUC, specificity, and sensitivity.

Table 5.4 in chapter 5 shows, the DCNN-MI-RF is the conqueror CAD model based on all metrics values. Based on this table, we see that both deep learning models can differentiate TPs and FPs more accurate than the hand-crafted feature-based method. The high dependency of the hand-crafted feature-based model on specific lesion-features may justify this issue for this model. In comparison the deep learning models are less influenced by the specific hand-crafted steps than the handcrafted feature-based framework, resulting in the better chance of detecting masses and the significantly better breast cancer detection performance by the DCNN-MI-RF. Moreover, I investigated cases of masses in the training set, which the DCNN MI-RF framework failed for the correct train. Images with the label of Ductal Carcinoma in Situ scored very low because in our training set we had a few cases with this label. Other samples were masses that appeared very close to the breast boundary that the system could not recognize them. The recognition of these cases could be improved if more training samples can be provided from real medical images or augmented data using generative adversarial neural networks [21]. Though DCNN- and DCaRM-based frameworks do not require the segmentation and feature extraction stages when compared to the hand-crafted, feature-based model, they needed big enough training data. Nevertheless, for further studies, we should collect a larger DBT dataset for training and test deep learning models and further investigation for application of the new augmentation techniques on DBT data like the generative adversarial neural networks algorithm for DBT data.

Table 6.1 presents a summary of selected studies in the literature with reported results. Note that we reviewed these models for mass detection in DBT methods in Chapter 2. Unfortunately, there is not any publicly available data for DBT to evaluate our CAD frameworks with several approaches. The models in the literature have obtained results from different private data, so a

TABLE 6.1: A comparison of selected studies in the detection of masses in the DBT datasets.

Author (Year)	Classifier	Size of DBT dataset	Performance
Fotin et al. [35] (2016)	DCNN	344	89% sensitivity 86.40% accuracy
Samala et al. [76] (2016)	DCNN	94	80% sensitivity over 0.80 AUC
Reiser et al. [70] (2006)	LDA	36	90% sensitivity
Kim et al. [47] (2016)	SVM	160	0.847 AUC
Schie et al. [91] (2013)	NN	752	80% sensitivity
Chan et al. [26] (2008)	LDA	100	80% sensitivity 0.91AUC
Palma et al. [65] (2014)	SVM	101	90% sensitivity
Wei et al. [95] (2011)	LDA	50	0.866 ACU
Morra et al. [61] (2015)	—	89	89% sensitivity
Chan et al. [24] (2005)	LDA	26	85% sensitivity 0.91AUC
DCNN MI-RF in Table 5.4 [104]	MI-RF	87	86.6% sensitivity 0.87 AUC 87.5 % specificity 86.81 % accuracy
DCaRBM MI-RF in Table5.4 [104]	MI-RF	87	81.8% sensitivity

quantitative comparison between our CAD systems and these models is not possible.

Among these studies, three works have developed deep learning based CAD frameworks Samala

et al. [76], Fotin et al. [35], and Kim et al. [47] that reported the impressive results using deep learning methods. These studies presented CAD frameworks for detection of malignant masses in DBT volumes based on both hand-crafted feature- and DCNN-based models. Samala et al. [76] proposed a DCNN-based CAD framework with the prescreening stage that provided the mass candidates to the DCNN with four convolutional layers and three fully connected layers. The proposed model was first trained on the mammography ROIs, then all the weights of the first three convolutional layers were frozen, and the rest of the network was trained using the DBT ROIs. The system in [76] was evaluated on 94 DBT volumes with 89 lesions and achieved 80% sensitivity at 2.94 false positives per volume. Fotin et al. [35] have used the DCNN system like the AlexNet model [51]. They trained the model on the generated ROIs containing $1{,}864$ lesions from mammography, and 339 lesions from slice-by-slice basis data of the DBT volumes. They obtained the 89% sensitivity at 3.25 false positives per volume 334 DBT volumes containing suspicious and malignant lesions. Kim et al. [47] proposed the DCNN model to identify the latent bilateral feature representations of masses in DBT volumes. They reported the AUC of 0.847 for the latent bilateral feature representation model.

Let us mention other works that have been developed based on integrated information from 2D slices of DBT images for mass detection in DBT volumes. Schie et al. [91] have proposed the framework based on a two-stage approach. First, information from 2D slices of DBT images has gathered using the 2D CAD framework [46]. Then the findings from 2D slices have merged by applying some rules on finding regions in different slices. They used the data from 192 patients where 49 patients were diagnosed resulting in 80% sensitivity at 3 false positives per volume. Reiser et al. [70] proposed the model that has detected masses individually within 2D projection views then recombined the detections in the 3D space. They reported a sensitivity 90% for 36 DBT volumes.

A fair comparison of our results and those results of these studies is not possible since the datasets and the measurement methods are not the same. The specific device provided each dataset with the different number of PVs and from patients with various clinical cases. For example, the dataset we have in hand includes real masses, vascular intersections, and other high-intensity irregular shaped objects but not all clinical cases. The current extracted features

for the hand-crafted feature-based models are intuitive for our dataset and is not applicable on the other datasets in other studies. For the deep learning models, some intrinsic factors from the dataset such as the number of training data and the heterogeneous data directly affect the design of the system. Moreover, each CAD system was designed to achieve the promising results on the dataset with the particular goal of the study. For instance, in this thesis, our goal was to reduce the false positive detection rate and the false negative rates. The consequence of non-public DBT data set is that we cannot even implement the published works in Table 6.1 on our dataset for comparing their results with our framework results. However, the DCNN MI-RF with best results among other frameworks in this study obtained 86.81% accuracy, and 86% sensitivity with an AUC of 0.87 for the DBT testing dataset with 28 DBT images. So based on our essential goal in this study that was to improve the sensitivity and the specificity rates of mass detection on our DBT dataset, using these impressive results, we can claim that our framework achieved this goal.

6.2 Conclusions

Digital breast tomosynthesis images as a breakthrough in mammography can provide a clearer, more accurate view compared to the digital mammography images. This imaging modality allows the radiologists to see through layers of breast tissues from different angles. A DBT image is a quasi 3D volume that is reconstructed from the limited number of low-dose projection views acquired over a limited angular around the breast.

Throughout this thesis, we addressed the challenges in developing new CAD systems for DBT images. We presented three frameworks for detecting spiculated masses in multi-layer breast tomosynthesis images based on hand-crafted feature, DCaRBM, and DCNN methods for extracting features from images and the MIL algorithms to classify DBTs. The most significant advantage of the proposed CAD frameworks is that the models learn complex patterns from 2D slices of DBT data (automatically or using hand-crafted features) and use all the information obtained from the slabs to analyze DBT images with weakly supervised classifiers based on multiple-instance learning algorithms. We used the separate MIL layer in deep learning based

CAD models as the same classification approach between all systems because the separate MIL layer in deep learning models allows us to have the fair comparison of the classification performance among three frameworks. Besides the advantage of the appropriate comparison, the separate MIL layer enables us to utilize different MIL models (MI-RF, MI-SVM, etc.) efficiently in our CAD frameworks. In this study, I showed that the DCNN MI-RF based CAD framework outperformed hand-crafted, feature-based, and DCaRBM-based models using the metrics of accuracy, AUC, specificity, and sensitivity.

Besides, we designed two other channels for detecting the micro-calcifications and bilateral asymmetry in images. For the bilateral asymmetry system, we provided the successful alignment of 2D slice images from the left and right breasts on DBT images. This approach overcame the breast registration challenge on the DBT data. There are limited studies on indication bilateral asymmetry on DBT images or even on mammogram images. So we can claim that this channel can serve as the first step for future studies on bilateral asymmetry detection in DBT images. Our micro-calcification detection framework also achieved the promising results for detecting clusters of micro-calcifications on images. However, our data for the micro-calcification detection channel was not sufficient to generalize the presented method to the population.

6.2.1 Future Directions

The methods used in CAD frameworks in this thesis based on Gabor filters, cooperative graph cut segmentation, TPS registration, deep CaRBM, DCNN, as well as MI-RF model, offer promising ways to detect spiculated masses and bilateral asymmetrical detection in DBT images. Although CAD frameworks obtained promising results in DBT classification, further improvements can be made when more DBT data are available. Lacking sufficient DBT data is the main drawback of this work. The efficiency and robustness of our developed CAD frameworks need to be investigated on the general DBT images of the population in the future. These CAD frameworks require various data with more micro-calcifications and architecture distortion samples to obtain satisfactory public performance, which can be used to detect all early signs of breast cancers. The proposed CAD frameworks in this study are not ready to be used for clinical applications;

many of their components need to be improved, and several directions of the research based on this study must be investigated.

- **Examine the Number of PVs**

  The number of projection views and tomographic angular range may affect the CAD systems; this effect should be examined in detail.

- **Examine Deep Learning Models using Transfer Learning on Mammogram Data**

  Available DBT data are insufficient in the world because the DBT image modality is the new X-ray image technique. One of the practical ways to examine our deep learning models is using the transfer learning using the large mammogram dataset. First, the deep learning model is trained on the large mammogram dataset includes various clinical cases. Then, the weights of the adjacent layers to the input layer are frozen, and the rest of the model is trained on the DBT data. This transfer learning approach may help to generalize deep learning models for analyzing the new clinical DBT images.

- **Examine Other Augmentation Techniques on DBT**

  In this study, we used some augmentation algorithms to increase the size of the DBT data artificially. For further studies, other augmentation techniques like the generative adversarial neural networks algorithm for DBT data should be investigated for the performance improvement of the deep learning models.

- **Improving Micro-calcifications Detection Approach**

  Two proposed robust deep learning approaches can be used to detect other early signs of breast cancers such as architectural distortion and micro-calcifications. Future work with more data will focus on combining 3D reconstructed imaging information with 2D information on adequate data. This combination may improve accuracy in other early

signs of breast cancer detection procedures. This improvement will be especially valuable when more clinical cases are used in the study.

- **Improving Bilateral Asymmetry Detection**

  The bilateral asymmetry detection method was not evaluated thoroughly. We designed the registration algorithm and region comparison algorithm in our framework and examined the system on the available data. But we did not know what asymmetric level will raise the alarm, given that we did not have image cases labeled in this category. So for further development of this system, we need to evaluate the system on the images with the malignant bilateral asymmetry label.

Bibliography

[1] Canadian cancer statistics. In *Canadian Cancer Statistics 2017. Toronto, ON: Canadian Cancer Society; 2017. Available at: cancer.ca/Canadian-Cancer- Statistics-2017-EN.pd*, 2017.

[2] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, and et.al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[3] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli. Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals. In *Computers in Biology and Medicine*, 2017.

[4] E. Amir, O. C. Freedman, B. Seruga, and D. G. Evans. Assessing women at high risk of breast cancer: a review of risk assessment models. In *JNCI: Journal of the National Cancer Institute*, volume 102, pages 680–691, 2010.

[5] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. In *Neural Computation*, volume 9, pages 1545–1588, 1997.

[6] F. J. Anscombe. The transformation of poisson, binomial and negative-binomial data. In *Biometrika*, volume 35, pages 246–254, 1948.

[7] C. Badea, K. Bliznakowa, Z. Kolitsi, and N. Pallikarakis. Noise removal in tomosynthestic mammographic imaging. In *in IFMBE Proc. 2nd European Medical and Biological Engineering Conference. BMBEC*, pages 924–5, 2002.

[8] J. Baker and Y. Lo. Breast tomosynthesis: state-of-the-art and review of the literature. In *Academic Radiology*, volume 18, pages 1298 – 1310, 2011.

[9] C. Balleyguier, J. Arfi-Rouche, L. Levy, P. R. Toubiana, F. Cohen-Scali, A. Y. Toledano, and B. Boyer. Improving digital breast tomosynthesis reading time: A pilot multi-reader, multi-case study using concurrent computer-aided detection (cad). In *European Journal of Radiology*, volume 97, pages 83 – 89, 2017.

[10] S. Bandyopadhyay. Breast asymmetry-tutorial review. In *International Journal of Computer Applications*, volume 9, pages 15–18, 2010.

[11] Y. Baneva, K. Bliznakova, L. Cockmartin, S. Marinov, I. Buliev, G. Mettivier, H. Bosmans, P. Russo, N. Marshall, and Z. Bliznakov. Evaluation of a breast software model for 2d and 3d x-ray imaging studies of the breast. volume 41, pages 78 – 86, 2017.

[12] A. S. Becker, M. Marcon, S. Ghafoor, M. C. Wurnig, T. Frauenfelder, and A. Boss. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. In *Investigative Radiology*, volume 52, pages 434–440, 2017.

[13] M. Berks, Z. Chen, S. Astley, and C. Taylor. Detecting and classifying linear structures in mammograms using random forests. In G. Székely and H. K. Hahn, editors, *Information Processing in Medical Imaging: 22nd International Conference*, pages 510–524, 2011.

[14] S. Bernard, S. Muller, and J. Onativia. Computer-aided microcalcification detection on digital breast tomosynthesis data: A preliminary evaluation. In *Digital Mammography*, pages 151–157, 2008.

[15] K. Bliznakova, Z. Bliznakov, and I. Buliev. Comparison of algorithms for out-of-plane artifacts removal in digital tomosynthesis reconstructions. In *Computer Methods and Programs in Biomedicine*, volume 107, pages 75 – 83, 2012.

[16] L. Bo, X. Ren, and D. Fox. Multipath sparse coding using hierarchical matching pursuit. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 660–667, 2013.

[17] F. L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 11,

pages 567–585, 1989.

[18] F. L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. In *Pattern Analysis and Machine Intelligence, IEEE Transactions.*, volume 11, pages 567–585, 1989.

[19] L. Breiman. Random forests. In *Machine Learning*, volume 45, pages 5–32, Oct 2001.

[20] L. M. Bruce and R. R. Adhami. Classifying mammographic mass shapes using the wavelet transform modulus-maxima method. volume 18, pages 1170–1177, 1999.

[21] F. Calimeri, A. Marzullo, C. Stamile, and G. Terracina. Biomedical data augmentation using generative adversarial neural networks. In A. Lintas, S. Rovetta, P. F. Verschure, and A. E. Villa, editors, *Artificial Neural Networks and Machine Learning – ICANN 2017*, pages 626–634, 2017.

[22] G. Carneiro, Y. Zheng, F. Xing, and L. Yang. Review of deep learning methods in mammography, cardiovascular, and microscopy image analysis. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing: Precision Medicine, High Performance and Large-Scale Datasets*, pages 11–32, 2017.

[23] M. A. Carreira-Perpinan and G. E. Hinton. On contrastive divergence learning. In *Artificial Intelligence and Statistics*, pages 33–40, 2005.

[24] H. Chan, J. Wei, E. Sahiner, B.and Rafferty, M. Wu, T.and Roubidoux, R. Moore, Kopans, D.B., L. Hadjiiski, and M. Helvie. Computer-aided detection system for breast masses on digital tomosynthesis mammograms: preliminary experience. In *Journal of Radiology*, volume 237, pages 1075–1080, 2005.

[25] H. Chan, J. Wei, Y. Zhang, M. Helvie, R. Moore, B. Sahiner, L. Hadjiiski, and D. Kopans. Computer-aided detection of masses in digital tomosynthesis mammography: comparison of three approaches. In *Medical Physics*, volume 35, pages 4087–4095, 2008.

[26] H.-P. Chan, J. Wei, Y. Zhang, M. A. Helvie, R. H. Moore, B. Sahiner, L. Hadjiiski, and D. B. Kopans. Computer-aided detection of masses in digital tomosynthesis mammography: Comparison of three approaches. In *Medical Physics*, volume 35, pages 4087–4095,

2008.

[27] S. Chang and N. Morgan. Robust cnn-based speech recognition with gabor filter kernels. In *Interspeech*, pages 905–909, 2014.

[28] S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, pages 537–546, 2008.

[29] R. S. de la Rosa, M. Lamard, G. Cazuguel, G. Coatrieux, M. Cozic, and G. Quellec. Multiple-instance learning for breast cancer detection in mammograms. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 7055–7058, 2015.

[30] F. S. S. de Oliveira, A. O. de Carvalho Filho, A. C. Silva, A. C. de Paiva, and M. Gattass. Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and svm. In *Computers in Biology and Medicine*, volume 57, pages 42–53, 2015.

[31] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. In *Artificial Intelligence*, volume 89, pages 31–71, 1997.

[32] G. Doran and S. Ray. A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. In *Machine Learning*, volume 97, pages 79–102, 2014.

[33] A. Elmoufidi, K. E. Fahssi, S. Jai-andaloussi, A. Sekkaki, Q. Gwenole, and M. Lamard. Anomaly classification in digital mammography based on multiple-instance learning. In *IET Image Processing*, volume 12, pages 320–328, 2018.

[34] S. S. J. Feng and I. Sechopoulos. Clinical digital breast tomosynthesis system: dosimetric characterization. In *Radiology*, volume 263, pages 35–42, 2012.

[35] S. V. Fotin, Y. Yin, H. Haldankar, J. W. Hoffmeister, and S. Periaswamy. Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and

deep learning approaches. In *Medical Imaging 2016: Computer-Aided Diagnosis*, page 97850X, 2016.

[36] I. J. Goodfellow, A. Courville, and Y. Bengio. Large-scale feature learning with spike-and-slab sparse coding. In *International Conference on Machine Learning (ICML)*, pages 1439—-1446, 2012.

[37] Y. Guo, R. Sivaramakrishna, C.-C. Lu, J. S. Suri, and S. Laxminarayan. Breast image registration techniques: a survey. In *Medical and Biological Engineering and Computing*, volume 44, pages 15–26, 2006.

[38] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[39] M. Heath, K. W. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer. The digital database for screening mammography. In *5th International Workshop on Digital Mammography, Toronto, Canada*, 2000.

[40] G. E. Hinton. Training products of experts by minimizing contrastive divergence. In *Neural Computation*, volume 14, pages 1771–1800, 2002.

[41] P. S. C. Ho. *Computer-aided detection and classification of microcalcifications in digital breast tomosynthesis*. PhD thesis, University of Oxford Doctor, 2012.

[42] S. Ioffe and C. Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning*, volume 37, pages 448–456, 2015.

[43] S. Jegelka and J. Bilmes. Submodularity beyond submodular energies: Coupling edges in graph cuts. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1897–1904, 2011.

[44] Z. Jiao, X. Gao, Y. Wang, and J. Li. A deep feature based framework for breast masses classification. In *Neurocomputing*, volume 197, pages 221 – 231, 2016.

[45] F. Jin, P. Fieguth, L. Winger, and E. Jernigan. Adaptive wiener filtering of noisy images and image sequences. In *International Conference on Image Processing*, volume 3, pages III–349–52 vol.2, 2003.

[46] M. Kallenberg and N. Karssemeijer. Computer-aided detection of masses in full-field digital mammography using screen-film mammograms for training. In *Physics in Medicine and Biology*, volume 53, page 6879, 2008.

[47] D. H. Kim, S. Kim, and Y. M. Ro. Latent feature representation with 3-d multi-view deep convolutional neural network for bilateral analysis in digital breast tomosynthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 927–931, 2016.

[48] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[49] P. Kohli, A. Osokin, and S. Jegelka. A principled deep random field model for image segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1971 – 1978, 2013.

[50] D. B. Kopans. Basic physics and doubts about relationship between mammographically determined tissue density and breast cancer risk. In *Radiology*, volume 246, pages 348–353, 2008.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012.

[52] D. S. Lalush, E. Quan, R. Rajaram, J. Zhang, J. Lu, and O. Zhou. Tomosynthesis reconstruction from multi-beam x-ray sources. In *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro*, pages 1180–1183, 2006.

[53] D. Laptev and J. M. Buhmann. Transformation-invariant convolutional jungles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3043–3051, 2015.

[54] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[55] C. Leistner, A. Saffari, and H. Bischof. Miforests: multiple-instance learning with randomized trees. In *Proceedings of the 11th European Conference on Computer Vision: Part VI*, pages 29–42, 2010.

[56] Y. Levakhina. Three-dimensional digital tomosynthesis: iterative reconstruction, artifact reduction and alternative acquisition geometry. Springer Fachmedien Wiesbaden, 2014.

[57] Y. Li, H. Chen, Y. Yang, L. Cheng, and L. Cao. A bilateral analysis scheme for false positive reduction in mammogram mass detection. In *Computers in Biology and Medicine*, volume 57, pages 84 – 95, 2015.

[58] Y. Lu, M. Yousefi, J. Ellenberger, R. H. Moore, D. B. Kopans, A. Krzyżak, and C. Y. Suen. 3d tomosynthesis to detect breast cancer. In *Handbook of Pattern Recognition and Computer Vision*, chapter 2.10, pages 371–393. World Scientific, 5th edition, 2015.

[59] M. Makitalo and A. Foi. Optimal inversion of the anscombe transformation in low-count poisson image denoising. In *IEEE Transactions on Image Processing*, volume 20, pages 99–109, 2011.

[60] T. Mertelmeier, J. Orman, W. Haerer, and M. Dudam. Optimizing filtered backprojection reconstruction for a breast tomosynthesis prototype device. In *Medical Imaging 2006: Physics of Medical Imaging*, volume 6142, pages 61420F–6140F–12, 2006.

[61] L. Morra, D. Sacchetto, M. Durando, S. Agliozzo, L. A. Carbonaro, S. Delsanto, B. Pesce, D. Persano, G. Mariscotti, V. Marra, P. Fonio, and A. Bert. Breast cancer: computer-aided detection with digital breast tomosynthesis. In *Radiology*, volume 277, pages 56–63, 2015.

[62] L. T. Niklason, B. T. Christian, L. E. Niklason, D. B. Kopans, D. E. Castleberry, B. H. Opsahl-Ong, C. E. Landberg, P. J. Slanetz, A. A. Giardino, R. Moore, D. Albagli, M. C. DeJule, P. F. Fitzgerald, D. F. Fobare, B. W. Giambattista, R. F. Kwasnick, J. Liu, S. J.

Lubowski, G. E. Possin, J. F. Richotte, C. Y. Wei, and R. F. Wirth. Digital tomosynthesis in breast imaging. In *Radiology*, volume 205, pages 399–406, 1997.

[63] L. Nyström, I. Andersson, N. Bjurstam, J. Frisell, B. Nordenskjöld, and L. E. Rutqvist. Long-term effects of mammography screening: updated overview of the swedish randomised trials. In *The LANCET*, volume 359, pages 909–919, 2002.

[64] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? In *Vision Research*, volume 37, pages 3311 – 3325, 1997.

[65] G. Palma, I. Bloch, and S. Muller. Detection of masses and architectural distortions in digital breast tomosynthesis images using fuzzy and a contrario approaches. In *Pattern Recognition*, volume 47, pages 2467 – 2480, 2014.

[66] S. Pereira, A. Pinto, V. Alves, and C. A. Silva. Brain tumor segmentation using convolutional neural networks in mri images. In *IEEE Transactions on Medical Imaging*, volume 35, pages 1240–1251, 2016.

[67] G. Peters, S. Muller, S. Bernard, R. Iordache, F. Wheeler, and I. Bloch. Reconstruction-independent 3d cad for calcification detection in digital breast tomosynthesis using fuzzy particles. In A. Sanfeliu and M. L. Cortés, editors, *Progress in Pattern Recognition, Image Analysis and Applications*, pages 400–408, 2005.

[68] G. Quellec, M. Lamard, M. Cozic, G. Coatrieux, and G. Cazuguel. Multiple-instance learning for anomaly detection in digital mammography. In *IEEE Transactions on Medical Imaging*, volume 35, pages 1604–1614, 2016.

[69] I. Reiser, R. Nishikawa, M. Giger, D. Kopans, E. Rafferty, T. Wu, and R. Moore. A multi-scale 3d radial gradient filter for computerized mass detection in digital tomosynthesis breast images. In *International Congress Series*, volume 1281, pages 1058 – 1062, 2005.

[70] I. Reiser, R. M. Nishikawa, M. L. Giger, T. Wu, E. A. Rafferty, R. Moore, and D. B. Kopans. Computerized mass detection for digital breast tomosynthesis directly from the projection images. In *Medical Physics*, volume 33, pages 482–491, 2006.

[71] I. Reiser, R. M. Nishikawa, A. V. Edwards, D. B. Kopans, R. A. Schmidt, J. Papaioannou, and R. H. Moore. Automated detection of microcalcification clusters for digital breast tomosynthesis using projection data only: A preliminary study. In *Medical Physics*, volume 35, pages 1486–1493, 2008.

[72] A. Rodríguez-Ruiz, M. Castillo, J. Garayoa, and M. Chevalier. Evaluation of the technical performance of three different commercial digital breast tomosynthesis systems in the clinical environment. In *Physica Medica*, volume 32, pages 767 – 777, 2016.

[73] B. Sahiner, H.-P. Chan, L. M. Hadjiiski, M. A. Helvie, J. Wei, C. Zhou, and Y. Lu. Computer-aided detection of clustered microcalcifications in digital breast tomosynthesis: A 3d approach. In *Medical Physics*, volume 39, pages 28–39, 2012.

[74] R. Salakhutdinov and G. Hinton. Deep boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 5, pages 448–455, 2009.

[75] R. K. Samala, H.-P. Chan, Y. Lu, L. Hadjiiski, J. Wei, B. Sahiner, and M. A. Helvie. Computer-aided detection of clustered microcalcifications in multiscale bilateral filtering regularized reconstructed digital breast tomosynthesis volume. In *Medical Physics*, volume 41, page 021901, 2014.

[76] R. K. Samala, H.-P. Chan, L. Hadjiiski, M. A. Helvie, J. Wei, and K. Cha. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. In *Medical Physics*, volume 43, pages 6654–6666, 2016.

[77] R. K. Samala, H.-P. Chan, L. M. Hadjiiski, K. Cha, and M. A. Helvie. Deep-learning convolution neural network for computer-aided detection of microcalcifications in digital breast tomosynthesis. In *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, pages 9785 – 9785 – 7, 2016.

[78] R. K. Samala, H.-P. Chan, L. M. Hadjiiski, and M. A. Helvie. Analysis of computer-aided detection techniques and signal characteristics for clustered microcalcifications on digital mammography and digital breast tomosynthesis. In *Physics in Medicine and Biology*, volume 61, pages 7092–7112, 2016.

[79] R. Schulz-Wendtland, M. Fuchsjäger, T. Wacker, and K.-P. Hermann. Digital mammography: An update. In *European Journal of Radiology*, volume 72, pages 258 – 265, 2009.

[80] I. Sechopoulos. A review of breast tomosynthesis. part ii. image reconstruction, processing and analysis, and advanced applications. In *Medical Physics*, volume 40, page 014302, 2013.

[81] I. Sechopoulos. A review of breast tomosynthesis. part i. the image acquisition process. In *Medical Physics*, volume 40, 2013.

[82] Shweta. 10 common type of cancer, 2017. Accessed at: `http://www.healthworry.com/10-common-type-of-cancer/`.

[83] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[84] J. Suckling, J. Parker, D. R. Dance, S. Astley, I. Hutt, C. R. M. Boggis, I. Ricketts, E. Stamakis, N. Cerneaz, S.-L. Kok, P. Taylor, D. Betal, and J. Savage. The mammographic image analysis society digital mammogram database. In *Digital Mammography: Proceedings of the 2nd International Workshop on Digital Mammography*, page 375–378, 1994.

[85] S. Suzuki, X. Zhang, N. Homma, K. Ichiji, N. Sugita, Y. Kawasumi, T. Ishibashi, and M. Yoshizawa. Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis. In *55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pages 1382–1386, 2016.

[86] K. Swersky, D. Tarlow, I. Sutskever, R. Salakhutdinov, R. Zemel, and A. Ryan. Cardinality restricted boltzmann machines. In *Neural Information Processing Systems*, volume 25, pages 3302–3310, 2012.

[87] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[88] D. Tarlow, I. Givoni, and R. Zemel. Hop-map: efficient message passing with high order potentials. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 812–819, 2010.

[89] B. S. Ted W Way, L. M. Hadjiiski, and H.-P. Chan. Effect of finite sample size on feature selection and classification: A simulation study. In *Medical Physics*, volume 37, pages 907—-920, 2010.

[90] R. Tibshirani. Regression shrinkage and selection via the lasso. In *Journal of the Royal Statistical Society. Series B (Methodological)*, volume 58, pages 267–288, 1996.

[91] G. van Schie, M. G. Wallis, K. Leifland, M. Danielsson, and N. Karssemeijer. Mass detection in reconstructed digital breast tomosynthesis volumes with a computer-aided detection system trained on 2d mammograms. In *Medical Physics*, volume 40, page 041902, 2013.

[92] S. Vedantham, A. Karellas, G. R. Vijayaraghavan, and D. B. Kopans. Digital breast tomosynthesis: state of the art. In *Radiology*, volume 277, pages 663–684, 2015.

[93] M. A. C. Vieira, P. R. Bakic, and A. D. A. Maidment. Effect of denoising on the quality of reconstructed images in digital breast tomosynthesis. In *SPIE Medical Imaging. International Society for Optics and Photonics*, volume 8668, pages 8668 – 8668 – 14, 2013.

[94] J. Wang. Hessian locally linear embedding. In *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*, pages 249–265, 2011.

[95] J. Wei, H.-P. Chan, B. Sahiner, L. M. Hadjiiski, M. A. Helvie, C. Zhou, and Y. Lu. Computer-aided detection of breast masses in digital breast tomosynthesis (dbt): improvement of false positive reduction by optimization of object segmentation. In *Medical Imaging 2011: Computer-Aided Diagnosis*, volume 7963, pages 7963 – 7963 – 6, 2011.

[96] J. Wei, H.-P. Chan, L. M. Hadjiiski, M. A. Helvie, g. Lu, C. Zhou, and R. Samala. Multichannel response analysis on 2d projection views for detection of clustered microcalcifications in digital breast tomosynthesis. In *Medical Physics*, volume 41, page 041913,

2014.

[97] R. Welikala, P. Foster, P. Whincup, A. Rudnicka, C. Owen, D. Strachan, and S. Barman. Automated arteriole and venule classification using deep learning for retinal images from the uk biobank cohort. In *Computers in Biology and Medicine*, volume 90, pages 23 – 32, 2017.

[98] Y. Xu, J. Y. Zhu, E. Chang, and Z. Tu. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In *Computer Vision and Pattern Recognition*, pages 964–971, 2012.

[99] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, and E. I. C. Chang. Deep learning of feature representation with multiple instance learning for medical image analysis. In *IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1626–1630, 2014.

[100] M. J. Yaffe. Mammographic density. measurement of mammographic density. In *Breast Cancer Research : BCR*, volume 10, page 209, 2008.

[101] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27*, pages 3320–3328, 2014.

[102] M. Yousefi, A. Krzyżak, and C. Suen. Convex cardinality restricted boltzmann machine and its application to pattern recognition. In *Advances in Artificial Intelligence*, volume 8436, pages 369–374, 2014.

[103] M. Yousefi, Y. Lu, C. Y. Suen, and A. Krzyżak. Computer aided diagnosis images to detect breast cancer. In *Computer Methods in Biomechanics and Biomedical Engineering (CMBBE), Symposium, Montreal, Canada*, 2015.

[104] M. Yousefi, A. Krzyżak, and C. Y. Suen. Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning. In *Computers in Biology and Medicine*, volume 96, pages 283 – 293, 2018.

[105] M. D. Zeiler. *Hierarchical convolutional deep learning in computer vision*. PhD thesis, New York University, 2014.

[106] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision*, pages 818–833, 2014.

[107] Y. Zheng, D. Liu, B. Georgescu, D. Xu, and D. Comaniciu. Deep learning based automatic segmentation of pathological kidney in ct: Local versus global image context. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing*, pages 241–255, 2017.

[108] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li. Multi-instance multi-label learning. In *Artificial Intelligence*, volume 176, pages 291–2320, 2012.

[109] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *Computing Research Repository*, volume arXiv:1612.05968, 2016.

Appendix A

Platforms and Tools

Operating Systems and Hardware

- Linux, and macOS operating systems
- Pycharm, and Cython IDE
- Two NVIDA GTX 1080 11 GB GPUs

Computer Languages and Libraries

- Python and C++ languages
- Python liberates such as PIL, scipy, scikit.
- Tensorflow Library
- Open CV
- Python wrapper for some C++ scripts