THE UNIVERSITY *of* ADELAIDE

Automated Detection, Segmentation and Classification of Masses from Mammograms using Deep Learning

by

Neeraj Dhungel

A thesis submitted in fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering, Computer and Mathematical Sciences
School of Computer Science

October 2016

Contents

List of Figures

Abstract

Breast cancer is considered to be one of the major contemporary problems affecting the lives of thousands of women worldwide. One of the most effective tools in the fight against this disease is early detection based on the manual analysis of X-ray mammograms. This manual process of interpretation of mammograms involves the detection of breast lesions (e.g., masses), the segmentation of lesions boundaries and the classification of lesions based on their shape, appearance and texture features. This manual analysis of breast lesions from mammograms presents large interpretation variability amongst radiologists. This variability can be reduced with the aid of computer aided diagnosis (CAD) systems that can act as a second reader in the analysis of breast lesions. However, for a CAD system to be useful in a clinical setting, it must effectively classify lesions as benign or malignant.

Detection, segmentation and classification of breast lesions are the main three steps involved in fully automated CAD systems that can work in the analysis of mammograms. Building a CAD system is difficult because mammograms are marred by low signal to noise ratio for the visualisation of breast lesions. In addition, breast lesions present a large variation in terms of shape, size and appearance. A large number of methods have been applied for building automated CAD systems for both types of lesions, namely mass and micro-calcification, but in this work we focus only on the analysis of masses. The major drawback of current approaches is that they generate a large number of false positives and miss a fair amount of true positive regions during the mass detection stage. Furthermore, mass segmentation is generally based on active contour models and graph-based approaches that rarely capture the large shape and appearance variations of breast masses. Finally, mass classification is generally implemented using sub-optimal hand-crafted features and machine learning classifiers such as support vector machines (SVM), linear discriminant analysis (LDA), artificial neural net (ANN), etc. One major limitation of the majority of existing CAD systems is that most of them require manual intervention to obtain mass candidates for segmentation and classification.

This thesis presents a new approach based on recently developed deep learning models to develop a fully automated CAD system for automated detection, segmentation and classification of masses from mammograms. Our proposed solution to the mass detection problem consists of three stages: 1) mass candidate generation using multi-scale deep learning and Gaussian mixture models, 2) false positive reduction with a cascade of deep learning and random forests classifiers, 3) candidate refinement with a local search algorithm based on Bayesian optimisation. Our proposed mas segmentation methods are based on two kinds of structured output learning methods, namely: 1) structured support vector machine for parameter estimation and graph cut for inferring the segmentation labels, and 2) truncated fitting for parameter learning and tree re-weighted belief propagation for inference. The resulting segmentation is then refined using an active contour model. Our proposed mass classification deep learning method is modelled with a two-step training procedure, where the first step is based on a pre-training stage that estimates a large set of hand-crafted features, which is followed by a fine-tuning step that learns a classifier (that classifies masses into benign and malignant). Finally, we integrate our mass detection, mass segmentation and mass classification methods into a fully automated CAD system for the analysis of masses in mammograms. We validate our methodology on two publicly available datasets (INbreast and DDSM-BCRP) using different performance measures such as average Dice index for segmentation, free receiver operating curve (FROC) and average precision curve for detection, receiver operating curve (ROC), area under curve (AUC) and accuracy for classification. The experiments show that our methodology for detection, segmentation and classification of breast masses achieves competitive results with respect to the current state-of-the-art techniques in terms of all performance measures mentioned above.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed: ______________________________

Date: ______________________________

Publications

My thesis is based on the content of the following peer-reviewed conference and journal papers:

- Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. Deep structured learning for mass segmentation from mammograms. In *IEEE International Conference on Image Processing (ICIP)*, 2015.
  (DOI: 10.1109/ICIP.2015.7351343)
  (This paper was selected among the top 10% of the accepted papers of the conference)

- Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. Tree RE-weighted belief propagation using deep learning potentials for mass segmentation from mammograms. *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2015.
  (DOI: 10.1109/ISBI.2015.7163983)

- Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. Deep learning and structured prediction for the segmentation of mass in mammograms. *18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.
  (DOI: 10.1007/978-3-319-24553-9_74)

- Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests. *International Conference on Digital Image Computing: Techniques and Applications (DICTA)* 2015.
  (DOI: 10.1109/DICTA.2015.7371234).
  This paper was first presented at 3rd workshop on Breast Image Analysis (BIA) as part of 18th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI-BIA), Munich, Germany, which gave us permission to publish this paper elsewhere.

- Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. The Automated Learning of Deep Features for Breast Mass Classification from Mammograms. *19th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016.
  (Accepted for Publication)
  MICCAI 2016 has awarded me with a student travel grant for this paper.

- Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. A Deep Learning approach to fully automated analysis of Masses in Mammograms. *Submitted to Medical Image Analysis (MedIA)* .
  (Under Review)

Acknowledgements

I would like to extend my sincere gratitude to all my family, friends and supervisors who stood behind me and encouraged during my PhD research.

More importantly, I would like to thank my principal supervisor Assoc. Prof. Gustavo Carneiro for his excellent guidance and mentorship that allowed me to complete my PhD in the given time-frame. He also gave me the opportunity to work on machine learning during my PhD despite knowing my limited background in that topic.

I would like to appreciate the help and supervision provided by Prof. Andrew P. Bradley (The University of Queensland), my co-supervisor Dr. Tat-Jun Chin and Dr. Zhi Lu (University of South Australia). I would also like to thank my friends Zhibin Lao, Gabrial Maicas and Dr. Tuan Anh Ngo (Vietnam National University of Agriculture), who have shared with me the knowledge as well as the experience in other areas of life.

I would also like to thank the graduate office of The University of Adelaide for providing me with the Adelaide Scholarship International (ASI) and the Australian Research Council's Discovery Projects funding scheme (project DP140102794) for providing me financial support to attend various international and national conferences. Finally, I also thank all the staff of the Australian Centre for Visual Technologies (ACVT), School of Computer Science for academic and ethical work environment.

Finally, I would also like to acknowledge the funding provided by Undergraduate Grant Commission (UGC), Nepal, when I was working as a faculty member at the Kathmandu University, Dhulikhel, Nepal for our project "Brain Tumour Segmentation", which motivated me to pursue the PhD degree in the field of medical image analysis at The University of Adelaide. I would like to thank my dearest friend Sujan Adhikari, who was also the team leader of this project (now pursuing PhD degree at Nanyang Technological University, Singapore) for his help and support.

August, 2016

Neeraj Dhungel

Dedicated to my family especially for my wife and friends for their unconditional love and support.

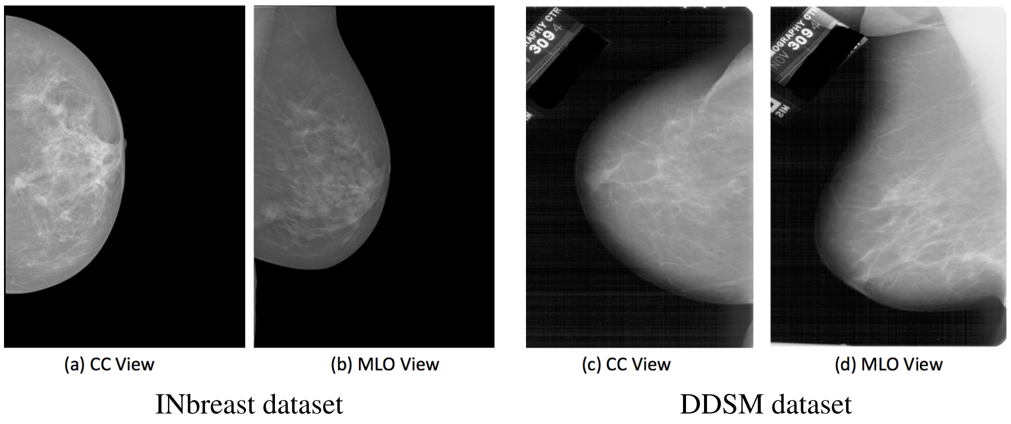
Chapter 1

Introduction

Breast cancer is regarded to be one of the major health issues worldwide. Nowadays, breast cancer accounts for 23% of all diagnosed cancers and 14% of cancer related deaths [1]. Over the past decade, the adoption of breast screening techniques based on X-ray mammography has contributed to the early detection of the disease, which has helped in the reduction of the mortality rate [2, 3] because women can receive proper treatment in the early stages of the disease.

Breast screening using X-ray mammograms is performed by taking images of the same breast from two different viewpoints, namely: mediolateral oblique (MLO) view and craniocaudal (CC) view, as shown in the Fig. 1.1. These images are then used for the detection and segmentation of breast lesions, such as masses and calcifications, in order to help radiologists assess the risk of breast cancer, where particular shape and appearance features of such lesions represent markers that the lesions are either malignant or benign. Masses are usually grey to white in pixel intensity, and geometrically they can have the oval, irregular or lobulated shapes with spiculated, circumscribed, obscured or ill defined margin [4, 5], whereas micro-calcifications are smaller rounded bright regions in the breast [4, 5] as shown in Fig. 1.2. In general, a breast mass is considered to be malignant if its shape is irregular or spiculated, and the classification of micro-calcifications are based on their size, shape, number and distribution [4, 5]. Although both types of lesions are important, in this thesis we focus only on the analysis of breast masses.

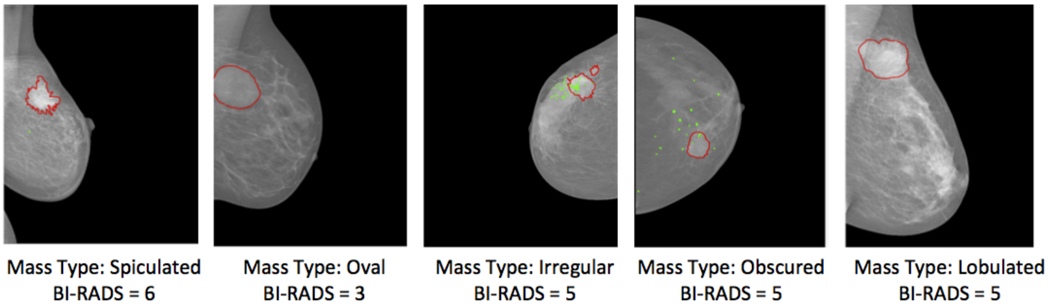
Detection, segmentation and classification of masses in mammograms is mostly done manually (see Fig. 1.3), which is a time consuming and subjective task that depends on the radiologist's expertise and fatigue level [4]. For instance, the sensitivity of the manual mass detection and classification fluctuates between 80% and 90% with specificity around 91% [6]. One way of increasing such specificity and sensitivity is with the double reading of mammograms, which has been found to increase sensitivity by 9% and to decrease the number of women recalled for further exams by 45% [7].



(a) CC View (b) MLO View (c) CC View (d) MLO View

INbreast dataset DDSM dataset

Figure 1.1: Some examples of mammogram images.



Mass Type: Spiculated BI-RADS = 6

Mass Type: Oval BI-RADS = 3

Mass Type: Irregular BI-RADS = 5

Mass Type: Obscured BI-RADS = 5

Mass Type: Lobulated BI-RADS = 5

Figure 1.2: Types of lesions in mammograms from the INbreast dataset, where we denote masses using red contours and calcifications using green contours.

Given the benefits of double reading in mammograms, a natural question that arises is if a CAD system can work as a second reader [8]. Evidently, a CAD system to be used as a second reader has to perform effectively in terms of having high sensitivity and specificity [8]. It has already been demonstrated by several studies [9–11] that radiologist's sensitivity in the detection and classification of masses improved by almost 10% with specificity at the same level with the use of a CAD system as second reader. A CAD system can be semi-automatic, requiring intervention by an expert at some stage, or fully-automatic, which requires no expert assistance. Similarly to the manual analysis of masses described above, a CAD system for the automated analysis of masses works in general in three steps: detection, segmentation and classification of masses. These three steps are challenging because of the variability of shape, size, appearance and location of masses in mammograms, low signal to noise ratio of the visual appearance of masses and lack of publicly available, precisely annotated datasets with full field digital mammograms (FFDM), which is currently the main imaging modality used in breast screening programs.

Over the years, a number of works have been proposed for the automated detection, segmentation and classification of masses from mammograms. Generally, mass detection from mammograms is carried out by detecting mass candidate regions, which is followed by false positive

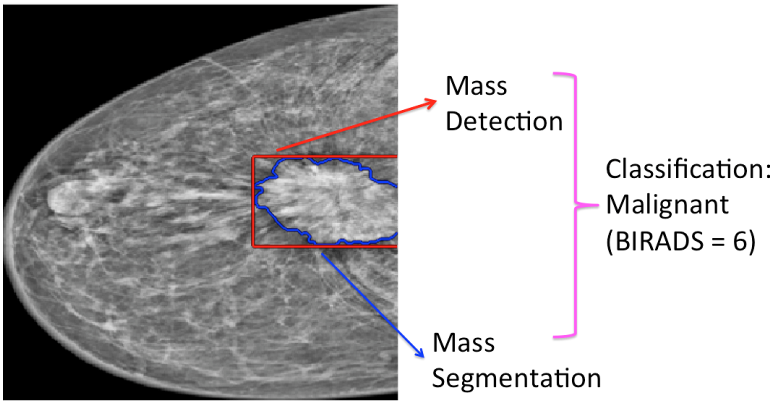


Figure 1.3: Mass detection, segmentation and classification.

reduction with different types of machine learning classifiers [4, 12]. Mass candidate detection is usually performed with techniques such as thresholding [13–17], edge detection using different types of filters [18–24], deformable models based on active contour models [25–30] and statistical methods such as region growing [31–33], region clustering using k-means [34–36], and Markov random field (MRF) [37–39]. In general, the mass candidate detection stage produces a high rate of false positives per image and requires another stage of false positive reduction with the extraction and classification of hand-crafted features [17, 40–46]. The main disadvantage of these systems is that they tend to produce high false positive rates [46] mostly because the mass candidate detection and hand-crafted features are sub-optimally designed to represent a human expert knowledge about the appearance and geometry of masses.

Mass classification into benign/malignant is a two-stage process where the first stage is the segmentation of mass and second stage is the mass classification. Segmentation is usually done in order to extract the geometrical features from the segmented mass contour. Traditional active contour [30, 42, 47–50] and graph based methods [51–55] are the two most successful algorithms for the mass segmentation. The main problem with these mass segmentation approaches is that they rely on manually defined shape and appearance terms and generally use a sub-optimal cross validation method to learn the parameters of the segmentation model. Mass classification is usually carried out with the extraction of hand-crafted features that are used by machine learning classifiers, such as linear discriminant analysis (LDA),artificial neural network (ANN), support vector machines (SVM), and etc [17, 40–45, 47, 56–58]. The major drawbacks of these mass classification methods is the sub-optimality in the design of features (similar to the problem of mass candidate detection) and these systems are generally semi-automated requiring the manual selection of mass candidates [48, 50, 59].

Deep learning models with its hierarchical feature representation have produced better classification accuracy compared to other machine learning techniques that use hand-crafted features [60–64]. Deep learning models are trained, such that the features automatically learned at each level of the model hierarchy are optimised in order to minimise a loss function based on detection, segmentation or classification results. This means that deep learning models can have features that are optimal for detection, segmentation or classification tasks, which is a tremendous improvement compared to the aforementioned hand-crafted features that are designed without minimising any loss function. However, training deep learning models is difficult because of their high capacity and the fact that in medical image analysis problems (e.g. breast mass analysis from mammograms), it is hard to find annotated datasets containing large amounts of training samples that would allow for a robust training. In spite of that, deep learning models have been explored for solving various medical image analysis problems, such as mitosis detection [65], lymph node detection, [66] and high-level classification of multi-modal input [67]. Thus, a natural question that arises is if deep learning models can perform better than the state-of-the-art methods in CAD systems that analyse mammograms. This question has been our main drive to test deep learning as the underlying framework for solving the problem of mass detection, segmentation and classification in mammograms.

1.1 Motivation

This thesis proposes a combination of several machine learning techniques and deep learning models [60, 61, 68, 69] for the detection, segmentation and classification of breast masses from mammograms. The primary motivation for using deep learning as our main underlying framework lies in its capability of learning a rich hierarchy of features by minimising a loss function that directly optimises the detection, segmentation and classification tasks. This is a great advantage over the use of hand-crafted features, which as explained above, is sub-optimal for the sought detection, segmentation and classification tasks. The motivation for combining other machine learning techniques with deep learning models lies in the need to regularise the training process given the large capacity of deep learning models, which can overfit the training set, particularly when this set does not contain large amounts of annotated samples. The fact that fully automated systems for mass detection, segmentation and classification in mammograms are rarely reported in the literature is another motivating factor for developing such system in this thesis.

Our mass detection methodology consists of several cascades of deep learning [60, 61, 68, 69] and random forest (RF) [70] classifiers, which is followed by a detection refinement process using a local search algorithm based on Bayesian optimisation [64]. The motivation of using the cascades of deep learning and random forests is the fact that the ability of deep learning method to reduce the false positive saturates after a certain number of cascade levels, so we extract hand-crafted features from the remaining candidates and input them to a cascade of RF classifiers that reduces the false positive rate per image. A recent study by Fernandez et al. [71] that shows that the performance of RF classifiers with hand-crafted features is better than other machine learning classifiers motivated us to use them for this final false positive reduction stage. In addition to this, we perform the detection refinement using a local search based on Bayesian optimisation [64], which has recently produced state-of-the art results in object localisation in computer vision.

We also propose a mass segmentation algorithm using a combination of several deep learning methods [60, 61, 68, 69] that form the unary potential functions in two different probabilistic graphical models [72–75] that solve a structured output learning problem. The primary motivation of combining probabilistic graphical models with deep learning models is that they have produced the current state-of-the-art results in semantic segmentation [62–64] and our method aims to produce similar results for the mass segmentation problem. The final segmentation is refined using an active contour model [76] as it helps to improve the segmentation accuracy.

Finally, we use a two-stage transfer learning approach for training our proposed mass classification system based on a deep learning model [60, 61], which uses the mass detection and segmentation results as its input. The first stage consists of pre-training the deep learning model, which is trained to regress the hand-crafted features and the second stage comprises a fine-tuning of the pre-trained model by minimising the classification loss. The extensive use of hand-crafted features in mass classification [36, 42, 44, 47, 48, 50, 52, 53] motivated us to pre-train our deep learning models with such features. Finally, we integrate our proposed mass detection, segmentation and classification methods to build the fully automated CAD system. We call our system "fully automated" because our proposed methodologies for mass segmentation and classification do not require any manual intervention during inference. However, it is important to note that in a real clinical setting, some manual intervention would be necessary to prune out the false positive detections. Furthermore, it is important to acknowledge that during the training stage, there is an amount of effort required to setup, tune and validate the model.

We test our methodologies on two datasets: INbreast [77] and DDSM-BCRP [78]. The motivation for using these two datasets is that they are available publicly and our results can be used as baseline by other researchers. The INbreast [77] dataset consists of 115 cases containing 410 images with all types of lesions including malignant, benign and normal, whereas the DDSM-BCRP [78] dataset contains 79 cases for training and 80 cases for testing containing only malignant lesions. The INbreast [77] dataset has accurately annotated full-field digital mammograms (FFDM), whereas DDSM-BCRP [78] contains digitised mammogram films with rough annotations [79].

1.2 Contributions of This Thesis

The main contributions of this thesis are as follows:

1. We develop two new methodologies for the segmentation of masses based on two types of structured output prediction models, which differ in the way they are trained and tested. One of the models uses truncated fitting [72] whereas the other uses the structured support vector machine (SSVM) [74, 75] for learning the parameters. Inference in one of the models is done with the tree re-weighted belief propagation (TRW) algorithm [72, 73] and the other model uses graph cuts [75, 80]. These two structured output prediction models use shape models from the following deep learning models: deep convolutional neural network (CNN) [60, 61] and deep belief network (DBN) [68, 69]. Both models produce state-of-the-art results in INbreast and DDSM-BCRP datasets. We explain our segmentation methodologies in Chapters 4, 5 and 6,

2. We present a novel machine learning approach for the detection of masses in mammogram, which combines cascades of CNN and RF classifiers. Our proposed mass detection method produces competitive results compared to the state-of-the-art work in mass detection [46]. Our methodology for mass detection is explained in Chapter 7,

3. We introduce a new classification methodology based on a two-step transfer learning approach using deep convolutional neural network (CNN) [60, 61], where the first step regresses the value of hand-crafted features, followed by a second step consisting of fine tuning based on breast mass classification [81]. Our proposed two-step transfer learning approach for mass classification produces better results compared to the state-of-the-art methods that use hand-crafted features [81] in automated and manual settings . We explain our classification methodology in Chapter 8,

4. Finally, we present a fully automated system for the detection, segmentation and classification of masses from mammograms. We add a refinement step to our mass detection methodology [46] using a local search based on Bayesian optimisation [64], which increases the precision of mass detection process. We also propose a segmentation refinement using an active contour model that increases the segmentation accuracy. The detail of our fully automated system for detection, segmentation and classification is explained in Chapter 9.

1.3 Thesis Outline

In Chapter 2, we review mass detection, segmentation and classification algorithms. Our proposed methodologies for the detection, segmentation and classification of masses from mammograms are explained in Chapter 3. We show the applications of our methodologies in Chapters 4, 5, 6, 7, 8 and 9 for the problem of mass detection, segmentation and classification in mammograms. We start with the problem of mass segmentation in Chapters 4, 5 and 6 using two structured output prediction models. In Chapter 7, we explain our mass detection methodology using cascades of CNN and RF. In Chapter 8, we introduce a mass classification methodology based on a two-stage transfer learning [60, 61]. In Chapter 9, we propose a fully automated system for mass detection, segmentation and classification, where mass detection is refined using a local search algorithm based on Bayesian optimisation and mass segmentation is refined using an active contour model. Finally, we conclude the thesis summarising our contributions and possible future work for mammogram analysis in Chapter. 10.

Chapter 2

Literature Review

A fully automated CAD system for the analysis of breast masses from mammograms involves three crucial steps: detection, segmentation and classification, which can be schematically represented in Fig. 2.1 [82]. In this chapter, we review the main techniques that are used for the detection, segmentation and classification of masses from mammograms. We start with mammogram pre-processing techniques in Sec. 2.1 followed by breast mass detection in Sec. 2.2, segmentation in Sec. 2.3, classification in Sec. 2.4 along with some commercially available CAD systems for analysis of mammograms in Sec. 2.5 so as to motivate our proposed methodology.

2.1 Mammogram Pre-processing

The goal of the pre-processing stage is to enhance the signal to noise ratio between masses and normal breast tissue structures in mammograms using image processing techniques, such as multi-scale wavelet transform, histogram equalisation and contrast limited adaptive histogram equalisation (CLAHE) [4, 12, 83, 84]. The pre-processing using wavelet transform is done by first transforming the mammogram in to the wavelet space by choosing a specific type of mother wavelet, where the resulting wavelet coefficients from the transformation are modified to enhance the mass features [4, 12]. Finally, the pre-processed mammogram using wavelet based methods is obtained by performing the inverse wavelet transformation [4, 12]. Similarly, we describe the other two widely used pre-processing techniques based on histogram equalisation for the analysis of mammogram [83, 85] in upcoming paragraphs.

Histogram equalisation [83] is a method for adjusting image intensities in order to enhance the contrast of an image. The objective of the histogram equalisation is to transform the image histogram to more uniform. Let us assume that we are given a grey value mammogram represented by $\mathbf{x} : \Omega \rightarrow \mathbb{R}$, with $\Omega \in \mathbb{R}^2$ denoting the image coordinate space. Let $\mathbf{x}(\mathbf{q})$ denote the grey

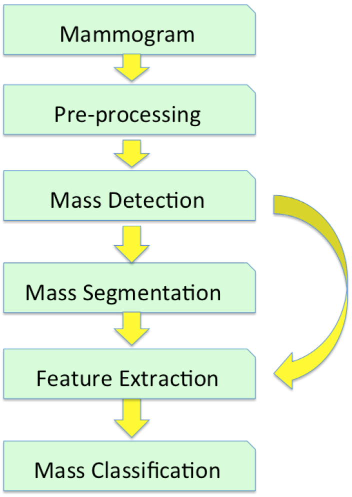


Figure 2.1: Main steps of a CAD system that analyses breast masses from mammograms.

level value of the pixel, quantised to be in the set $\mathcal{M} = \{0, 1, ..., L\}$, at the image grid location $\mathbf{q} \in \mathbb{R}^2$. Let n_i be the number of pixels at level $i \in \mathcal{M}$, such that the total number of pixels is $N = \sum_{i=0}^{L} n_i$. The histogram equalised image $\mathbf{x}_{\text{hist}}$ at grid location $\mathbf{q}$ can be represented as:

$$\mathbf{x}_{\text{hist}}(\mathbf{q}) = T(\mathbf{x}(\mathbf{q}); \theta_{\text{hist}}), \tag{2.1}$$

where $T(.)$ is the transformation mapping, θ_{hist} is the size of the bins for histogram and this transformation function is a monotonically increasing function in the interval $0 \leq i \leq L - 1$ such that $0 \leq T(.) \leq L - 1$. The transformation function $T(\mathbf{x}(\mathbf{q}); \theta_{\text{hist}})$ is defined in terms of cumulative distribution function (CDF) as:

$$\mathbf{x}_{\text{hist}}(\mathbf{q}) = \lfloor (L - 1) \sum_{i=0}^{\mathbf{x}(\mathbf{q})} p_i(\mathbf{x}) \rfloor, \tag{2.2}$$

where $p_i = n_i/N$ is the probability of occurrence of intensity level i in image $\mathbf{x}$, and $\lfloor (.) \rfloor$ rounds the value down to the nearest integer.

The transformation function which defines the histogram equalisation is based on intensity distribution of the entire image and is suitable for the global enhancement of the image. The local regions in the images may not be enhanced meaningfully by such histogram equalisation as these regions may have negligible impact on the computation of a global transformation function. Adaptive histogram equalisation [84] alleviates this problem by applying the operation in the neighbourhood regions surrounding the pixel in the image grid instead of using the whole image. Because adaptive histogram equalisation works over a local region, the contrast of the local region is enhanced, but this can also amplify the noise in that local region. Contrast limited adaptive histogram enhancement (CLAHE) [85] improves over the adaptive histogram equalisation by clipping the histogram at a certain threshold depending upon the size of the local region and normalisation. Variations of CLAHE have also been used for the detection of masses in dense breast [86] and as a preprocessing step for mass segmentation [30]. We also use CLAHE as a pre-processing step for our mass detection, segmentation and classification methodologies that we describe in Chapters 4, 5, 6, 7, 8.

2.2 Breast Mass Detection

2.2.1 Problem Definition

A breast mass is defined by a lump within the breast tissue, which may be benign (e.g. fibroadenoma, fibrocystic disease, breast abscess, and fat necrosis) or malignant (cancer). Breast masses are usually characterised by their geometrical location, shape and margin characteristics, and have brighter intensity compared to the surrounding breast tissue. The automated detection of masses in mammogram is challenging because of the large variation in terms of their geometrical structure (e.g. location, shape and margin characteristics) and low signal to noise ratio in relation to the surrounding breast tissues. Breast mass detection is the first stage (after preprocessing) of a CAD system for classifying suspicious breast masses; as a result, the accuracy of the CAD system depends upon a high sensitivity and specificity this detection step. However, current methodologies for mass detection still have high false positive rate per image. We believe that the main reason for that high false positive rate per image is that these methodologies use hand-crafted features that have not been optimally designed for breast mass detection. In addition to this, the ensuing segmentation stage usually requires the breast mass to be precisely detected in terms of position and scale. Below, we review different approaches for the detection of the breast masses in mammograms, describe their advantages and disadvantages, and motivate our methodology for mass detection that is described in Chapters 3 and 7.

2.2.2 Background

There are two main strategies for detecting masses in mammograms, namely: 1) mass detection using a single view image and 2) mass detection using multiple view images (CC and MLO) [4, 11, 82]. Both strategies work in two stages: the first stage consists of the detection of a high number of the mass candidates and the second stage comprises of the removal of false positives. Mass detection using multiple views is different based on the fact that this multi-view approach fuses information from both views to remove the false positives detected during the first stage [4, 11, 82]. Below, we start with the explanation of mass candidate detection techniques that are used in both the single and multiple view systems and then proceed to explain various techniques for false positive reduction.

The first stage of mass candidate detection is carried out with standard techniques, such as thresholding [13–17], gradient-based image segmentation [18–24], deformable models based on active contours [25–30], statistical methods such as region growing [31–33], region clustering with k-means [34–36], and Markov random field (MRF) [37–39]. Mass candidate detection based on thresholding, can rely on global [13–15] or local methods [16, 17]. The major drawback of thresholding methods is that they are not able to capture all the variations of intensity, texture and structure of masses in mammograms [4]. Gradient based methods such as Laplacian of Gaussian (LoG) [87, 88] and difference of Gaussian (DoG) [87, 88] are used for finding image regions in mammograms [18–23, 89]. The main issue with such methods is that the majority of regions found consists of regions that are visually salient in the sense that they are either brighter or darker than their neighbours. This is a necessary, but not a sufficient condition for a breast mass. Consequently, gradient-based methods are often followed by rounds of false positive reduction using machine learning classifiers.

Mass candidate detection using techniques, such as region growing [31–33], region clustering with k-means, [34–36] and Markov random field (MRF) [37, 38] use the statistical properties of the pixels and their neighbours in the image grid [31–33, 37, 38]. Region growing iteratively aggregates the pixels that have similar appearance characteristics [31–33] starting from a set of seed points in an image. Therefore, region growing requires a good selection of these seeds points, which is generally a hard task. The k-means clustering separates one or more regions based on the mean and variance of the intensity values of the pixels [34–36]. This characterisation based only on local appearance is generally too limited to represent the true complexity of a breast mass, and for this reason k-means clustering approaches tend to produce a large number of false positive candidates. Markov random field (MRF) introduces a spatial prior based on the idea that pixels that represents mass are likely to be clustered around a compact image region that have similar appearance. The main problem with MRF models is their large running time complexity, which makes the analysis of high-resolution images a hard task.

Deformable models, such as level set methods, and active contours are based on an energy minimisation problem [25–30] that depends on internal forces such as shape and curvature, and external forces such as image gradient for the mass candidate detection. The major problems with deformable models are that they require good initialisation (i.e., the initial contour must be close enough to the true mass candidates) and they also need an appropriate selection of a good set of weights for the energy terms in order to produce accurate detection results [39]. Template matching techniques [90–92] aim to match a mass template with candidates using simple matching criteria, such as least square distance or cross correlation. Template matching is also susceptible to high false positive rates and requires rounds of false positive reduction using machine learning classifiers [12].

Mass candidate detection is usually followed by a false positive reduction using different types of machine learning classifiers such as linear discriminant analysis (LDA), artificial neural network (ANN) and support vector machines (SVM) [4, 35, 44, 48, 50, 59, 82, 93, 94]. Most of the existing methodologies for false positive reduction using machine learning classifiers also employ feature extraction and selection [4, 35, 44, 48, 50, 59, 82, 93, 94]. The features that are used for this false positive reduction are usually hand-crafted and can be divided into morphological features, intensity features and texture features [4, 35, 44, 48, 50, 93]. Morphological features are computed based on the segmentation contour and they usually represent the geometrical properties of segmentation, such as margin spiculation and sharpness, area, circularity, rectangularity, perimeter, perimeter to area ratio, and normalised area length (NRL) features, such as boundary roughness, mean, entropy, standard deviation and area ratio zero crossing count. Mean contrast feature is computed as the ratio of grey scale values inside and in the vicinity of the segmented object. In addition to morphological and intensity features, global and local texture features are computed based on spatial grey level dependence (SGLD) matrix. SGLD matrix is defined as a distribution of occurrence of pixel intensity i with respect to a pixel intensity j, which varies with inter-pixel separation and direction. The local and global texture features that are based on SGLD matrix are energy, correlation, inertia, entropy, difference of moment, inverse difference of moment, sum average, sum entropy, difference entropy, sum variance, difference variance, difference average, information measure of correlation. An important point to note here is the fact that these features have been hand-crafted, which means that they cannot operate optimally for this classification of breast mass candidates. Consequently, even after this false positive reduction stage, it is likely that there are still false positives present in the set of mass candidates.

Mass detection using the multiple mammographic views are common practice in clinical environment. Radiologists often compare the asymmetry in terms of breast tissue by considering the density, size and shape differences between the left and right breasts, between different views of the same breast (CC and MLO), and also from mammograms of same patient over time. Bilateral subtraction measures the difference from normal symmetry between the left and right mammograms by subtracting the left and right mammograms at various intensity levels [95, 96] . Although bilateral subtraction is simple and does not produce a large number of false positives, it requires accurate registration between different mammograms (CC vs MLO, left vs right breast, and over time), which is a difficult task [12]. The multi-view approach using the CC and MLO images of the same patient usually starts by detecting the masses in each view independently and then comparing pair of masses from two different views. The distance of the mass candidates in the CC and MLO views from the nipple position in polar coordinate system is used as a measure to register the true positive regions that can be used to filter out false positives [97, 98]. Velikova et al. [99] proposed a Bayesian framework that estimates the links between the lesions detected in CC and MLO views. More recently, Amit et al. [100] proposed a frame work that combines unsupervised thresholding to generate mass candidates in both CC and MLO views and then estimate the correspondence between such candidates using hand-crafted features and a random forest [70] classifier. Radiologists have used mammograms of the same breast over a period of time to evaluate how suspicious lesions evolve within a specific timeframe. The detection of suspicious lesions within the particular timeframe is performed by identifying potential control points, such as junctions of curvilinear structures generated by the tissue, vessels and ducts, and estimating the correspondence between these landmarks in temporal images [101–103].

Recently, deep learning methods, such as deep convolutional neural nets (CNN), have produced state-of-the-art results in object detection [63, 64], and a natural question that arises in the context of mammogram analysis is if such methods can be applied for detecting breast mass candidates. Deep learning models automatically learn a complex hierarchy of features, eliminating the process of hand-crafting features. The main issue with the use of deep learning models in the analysis of mammograms is the fact that these models have quite large capacities, which means that they need large amounts of annotated training images in order to produce robust classifiers. Unfortunately, such large annotated training sets are usually unavailable for the problem of mammogram analysis, and one of the main challenges in medical image analysis is how to adapt deep learning models in this adversarial scenario. In this thesis, we propose a combination of several deep learning models using a cascade classifier [46, 104] for the problem of breast mass detection [46]. For the final false positive reduction, we rely on hand-crafted features and a random forest classifier [46, 70].

2.3 Breast Mass Segmentation

2.3.1 Problem Definition

Breast mass segmentation is the stage that commonly follows the mass detection step. The importance of this stage lies in the need to produce geometric features that characterise the shape, size, and boundary of a breast mass [11]. There exists numerous breast mass segmentation techniques, but CAD systems that depend on accurate breast mass segmentation methods, are rarely used in clinical practice because of the relatively low segmentation accuracy [105].

One of the reasons for this low segmentation accuracy is the reliance on traditional image processing and segmentation models, such as active contour models, based on the energy terms that do not characterise well the range of possible segmentation samples and the non-convex cost functions which produce sub-optimal segmentation results [39]. Another related problem with most of the existing breast mass segmentation methodologies is that they are generally semi-automatic, requiring radiologists to provide a region of interest (ROI). Moreover, most of these models are tested with private datasets that do not allow competing methods to be fairly compared [39, 106, 107]. In this section, we discuss the advantages and disadvantages of existing methods that are used for breast mass segmentation in order to motivate our proposed segmentation methodologies described in Chapters 4, 5 and 6.

2.3.2 Background

The majority of methodologies developed for the problem of segmenting breast masses are based on statistical thresholding, dynamic programming models, morphological operators, and active contour models. A statistical thresholding method, proposed by Catarious et al. [108], distinguishes pixels inside the mass area from those outside with an iterative thresholding algorithm, based on Fisher's linear discriminant analysis (FLDA). Even though this algorithm is successful to some extent, its major disadvantage is that it is prone to over-segmentation, which means that it classifies false positive pixels as true positive pixels and it is not robust to imaging conditions [30].

Song et al. [53] improved the model based on statistical thresholding [108] for breast mass segmentation with the use of dynamic programming based on hand-crafted shape and appearance models. The shape model is based on edge gradient, whereas appearance model is based on image grey values and the segmentation is found by estimating the minimum cut of a graph representing the image using dynamic programming. Similar graph-based methods have been explored by Timp et al. [54], Dominguez et al. [52] and Yu et al. [55], but with the use of various kinds of hand-crafted shape and appearance models. Compared to these methods, the main advantage of our approach is that our model automatically learns the shape and appearance features for this segmentation problem.

Morphological operators, such as the watershed method [109] and region growing [33, 110, 111] have been used for the breast mass segmentation problem. Watershed segmentation works by simulating the flooding process by considering the grey level images as topographic reliefs, where each relief is flooded from its minima and a dam is built when the two lakes meet. The set of all dams is regarded as watershed and in image segmentation, these dams represent the closed contours of the segmentation. Although, methods based on watershed segmentation are computationally very efficient, they suffer from over-segmentation, requiring post-processing by other techniques such as active contours [4, 112]. Similarly, region growing is an iterative method that assembles the pixels that have similar characteristics [33, 111, 113] given the set of seed points in an image. However, region growing is limited in providing sufficiently accurate segmentation because they only use semi-local grey level distributions without taking higher level information (e.g., shape model) into account.

Active contour models are one of the most explored methodologies for breast mass segmentation [42, 47–50], where the model proposed by Rahmati et al. [30] produces the state-of-the-art breast mass segmentation results. Rahmati et al.'s model is a level set method based on the maximum likelihood segmentation without edges that is particularly well suited to noisy images with weak boundaries. The main disadvantage of this method is that it is based on the minimisation of non-convex energy function that requires a good initialisation for the inference process. Moreover, the weights of the terms forming the energy function of the active contour models are usually arbitrarily defined, or estimated via a cross-validation process that usually do not produce an optimal estimation of these weights.

Deep learning models have produced state-of-the-art results in the field of semantic segmentation in computer vision [62, 114, 115]. Therefore, it is expected that such methods can replicate such outstanding results for the problem of breast mass segmentation from mammograms. However, these deep learning models face the same problems as described above for the problem of mass detection. In general, these approaches use a fully connected layer at the last layer of the CNN, which means that this layer has the same number of nodes as the input size, and these models can perform fast inference [62, 114, 115]. However, it has been observed that CNNs with such fully connected layer are unable to capture the fine details of the segmentation contour [114, 115]. Therefore, these CNNs are combined with other machine learning approaches, such as conditional random field (CRF), to make fine adjustments to edge boundaries [114]. Similarly defined deep learning models have been used in medical image segmentation problems, such as pancreas segmentation [116] and left-ventrical segmentation [117]. We alleviate the problem of having a limited number of annotated training samples by using a pre-trained model [67], by artificially increasing a number of training data [39, 67, 116], or by combining the deep learning model with other machine learning techniques [39, 117]. We describe our automated mass segmentation methodologies in Chapters 4, 5, 6, and 9.

2.4 Breast Mass Classification

2.4.1 Problem Definition

The final stage of the analysis process consists of classifying the breast masses into malignant or benign.

Breast mass classification systems presently depend upon hand-crafted features extracted from breast mass candidate regions and segmentations that are used by traditional machine learning classifiers in such classification [36, 42, 44, 47, 48, 50, 52, 53]. The main issue with these approaches is the process of hand-crafting features, which is sub-optimal given that such features are produced based on experts' biases. The other issue with these classification techniques is that mass candidates are usually selected manually [36, 42, 47, 48, 50, 52, 53], which is not acceptable if one aims to produce a fully automatic method. In the following section, we review current methodologies in order to motivate our mass classification system.

2.4.2 Background

The majority of current classification methods still relies on the manual localisation of masses as the automated mass detection is considered to be a challenging problem [44]. In addition, this classification usually depends on hand-crafted features, extracted from the detected image patches and their segmentation map, which are fed into the classifiers that try to classify masses into benign or malignant [44, 45, 47]. Such hand-crafted features include morphological features, intensity features and texture features (already discussed for breast mass detection in Sec. 2.2). These features are then used in a classification process based on traditional machine learning classifiers such as support vector machines (SVM), linear discriminant analysis (LDA) or multi-layer perceptron (MLP) [36, 42, 44, 47, 48, 50, 52, 53]. These methods also depend on feature extraction and selection processes, which are sub-optimal and can produce inaccurate classification results. Mass classification problems also present the issue of having limited availability of datasets, but it is important to notice that the INbreast dataset [77], which is publicly available, has been used to alleviate this issue [45].

Deep learning models have produced more accurate classification results compared to models based on hand-crafted features and traditional machine learning classifiers [61, 62]. One of the reasons behind the superior performance of deep learning models lies in their ability to automatically learn hierarchical features by minimising a classification loss function [67]. Our proposed methodology for breast mass classification uses deep convolutional neural network (CNN), where we apply a two-stage training process with the first stage comprising a pre-training step that regresses the values of hand-crafted features commonly used for breast mass classification. Our pre-training stage not only acknowledges the importance of these hand-crafted features, but also regularises the deep learning model training. The second stage is the actual breast mass classification, where we fine tune the pre-trained model from above with the minimisation of a classification loss. We explain our mass classification system in Chapters 3 and 8.

2.5 Mammogram Analysis Systems

According to Giger et al. [11], CAD systems can be broadly divided into two categories: computer aided detection (CADe) and computer aided diagnosis (CADx). CADe systems are mainly used for the detection of lesions and they help radiologists locate the suspicious regions in the breast area, whereas CADx systems focus on the classification (into malignant or benign) of suspicious lesions located by radiologists. CADx systems also assign a probability of malignancy of suspicious regions, leaving the final decision for radiologists. The risk of developing breast cancer is assigned in terms of Breast Imaging-Reporting and Data System (BI-RADS) [118], which ranges from 0 to 6, where BI-RADS scores between 0 and 3 are generally considered to be benign and BI-RADS scores between 4 and 6 are considered to be malignant.

The first reported CADe system was developed by Giger et al. [11, 82] for analysing mammograms. This system scanned film mammograms that were analysed in order to output the location of suspicious lesions (e.g., masses and micro-calcifications). In 1998, the first commercial CADe system was approved by the Food and Drug Administration (FDA), and since then several manufacturers such as R2 Technology, iCAD Medical Systems and Kodak have been working on the development of commercial breast CAD systems.

The latest version of commercial CADe systems report a mass detection sensitivity of 90% at a false positive rate of two candidates per case [119, 120], which means that current CADe systems exhibit comparable sensitivity with respect to experienced radiologists. However, the specificity of CADe system is far lower than that of experienced radiologists [82], and current CADe systems fail to provide the quality and accurate marks (in terms of localisation) as provided by experienced radiologists in the detection of breast masses [120].

The earliest CADx system were based on an expert system [11, 121, 122] for retrieving similar cases of a given lesion from a dataset containing malignant and benign cases. Giger et al. [11, 82, 123] developed a CADx system that could output a probability estimate of whether a suspected lesion, provided by a radiologist, was benign or malignant, and displayed the hand-crafted features based on the geometrical description of the lesions. In addition to this, this system could retrieve and display a specific number of similar cases for a given lesion. There have been independent studies [124–127] that have shown the benefits of using CADx systems. More specifically, these studies have shown that the performance of radiologists is significantly improved with the use of CADx systems. CADx systems have also been shown to reduce the variability of interpretation amongst radiologists. However, most of these CADx systems use hand-crafted features, which is sub-optimal, as explained above. Another issue with these CADx system is that they are not usually fully automated. In this thesis, we develop a fully automated CAD system for the detection, segmentation and classification of masses from mammograms. We address the issues mentioned above with the use of a cascade of deep learning and RF classifiers that improves the accuracy of mass detection results. We also address the problem of manual mass candidate selection, present in current CADx systems, by refining mass detection with a local search algorithm based on Bayesian optimisation. We also fix the problem of sub-optimality present in the hand-crafted features with the use of automatically learned features from deep learning models in detection and classification. Our fully automated CAD system detects 90% of masses at one false positive per image, has a segmentation accuracy of around 0.85 (Dice index), and classifies masses as malignant or benign with sensitivity (Se) of 0.98 and specificity (Sp) of 0.7.

2.6 Conclusions

The methodologies described above have been successfully applied for the automated detection, segmentation and classification of masses in mammogram. However, each methodology has some advantages and disadvantages, as discussed above. In a nutshell, mass detection models in mammogram produce high false positive rate per image [11] due to the low capacity of these models that are not capable of modelling the shape, size and intensity variations of the masses robustly. The main issue with the majority of mass segmentation methods is that they are based on methodologies that use sub-optimal training and inference algorithms. Similarly, current mass classification methodologies rely on hand-crafted features that are not optimised for breast mass classification. Another issue with current mass classification methodologies is that they usually depend on manually selected ROIs. In addition to this, most of the current detection, segmentation and classification methodologies are tested in private datasets, which make the comparison of such methodologies a difficult task.

Our proposed methodologies for mass detection, segmentation and classification addresses the issues outlined above by combining various machine learning techniques with deep learning models using publicly available mammogram datasets. This combination of deep learning with other machine learning approaches produces comparable or more accurate results than current state-of-the-art models for the detection, segmentation and classification of breast masses in these datasets. We show the results of our segmentation methodologies which use a combination of deep learning models and structured output prediction models in Chapters 4, 5 and 6. In Chapter 7, we show the results of automated mass detection using a cascade of deep learning and random forests classifiers. The results of our classification algorithm, which uses a transfer learning approach with a deep learning model, is shown in Chapter 8. Finally, we show the results of our proposed fully-automated CAD system in Chapter. 9.

Chapter 3

Methodology

3.1 Overview of the method

We have devoted this chapter for the explanation of the techniques that we use for the automated detection, segmentation and classification of masses in mammograms. Our methodologies are based on the combination of several machine learning techniques. We propose a cascade of deep learning [63] and random forest [70] classifiers for the detection of masses and Bayesian optimisation [64, 128] for the refinement of the masses' location and scale. For the problem of breast mass segmentation, we propose two structured output learning models [72–74] that use several deep learning shape models as their potential functions [60, 61]. This segmentation is then refined using an active contour model [29, 76]. Our mass classification approach is based on a transfer learning approach in which we pre-train our deep learning model [60, 61] using hand-crafted features. We then fine-tune this pre-trained deep learning model for breast mass classification. In the following sections, we describe these methods in detail.

3.2 Dataset Definiton

Let us represent the annotated dataset by $\mathcal{D} = \{(\mathbf{x}, \mathcal{A})_i\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x} : \Omega \to \mathbb{R}$ with $\Omega \in \mathbb{R}^2$ represents a mammogram, and $\mathcal{A}_i = \{(\mathbf{d}, \mathbf{y}, c)_j\}_{j=1}^{|\mathcal{A}_i|}$ denotes the annotation of $\mathcal{A}_i$ masses for mammogram i, where $\mathbf{d}(i)_j = [x, y, w, h] \in \mathbb{R}^4$ represents the left-top position (x, y) with width w and height h of the bounding box of the j^{th} mass of the i^{th} mammogram. Similarly, $\mathbf{y}(i)_j : \Omega \to \{0, 1\}$ denotes the segmentation map of the mass within the mammogram defined by the bounding box $\mathbf{d}(i)_j$, and $c(i)_j \in \{0, 1\}$ represents the class label of the mass that can be either benign, i.e., BI-RADS $\in \{1, 2, 3\}$ or malignant, i.e., BI-RADS $\in \{4, 5, 6\}$.

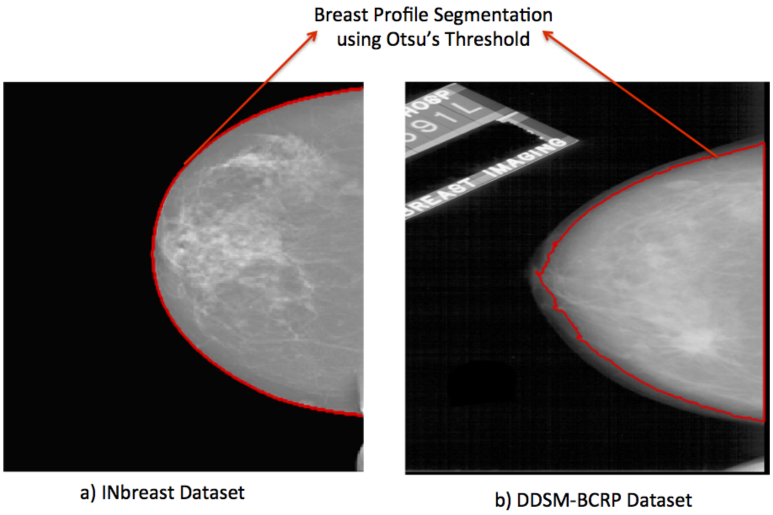


Figure 3.1: Breast profile segmentation using Otsu's thresholding.

3.3 Image Enhancement

3.3.1 Otsu's Thresholding

Otsu's thresholding [129] uses the intensity of the image for separating the object from the background. Assume that the mammogram $\mathbf{x}$ has quantised grey values ranging from 0 to L in $\mathcal{M} = \{0, 1, .., L\}$, and n_k represents the number of pixels at level $k \in \mathcal{M}$ such that the total number of pixels $N = \sum_{k=0}^{L} n_k$. Let $\mathcal{C}_1 = \{1, 2, ..., m\}$ and $\mathcal{C}_2 = \{m+1, m+2, ..., L\}$ represent the grey values of the foreground and background, respectively using a threshold m. The probability distribution of pixels at a grey level k is $p_k = \frac{n_k}{N}$, where $p_k \geq 0$ and $\sum_{k=1}^{L} p_k = 1$. The probability of occurrence of two classes (ω_1, ω_2) with threshold m is:

$$\omega_1(m) = \sum_{k=1}^{m} p_k \text{ and } \omega_2(m) = \sum_{m+1}^{L} p_k. \tag{3.1}$$

Similarly, we can define the class variance σ_1 of foreground and class variance σ_2 of background as:

$$\sigma_1(m) = \sum_{k=1}^{m} \frac{(k - \mu_1(m))^2 p_k}{\omega_1(m)}, \; \sigma_2(m) = \sum_{k=m+1}^{L} \frac{(k - \mu_2(m))^2 p_k}{\omega_2(m)}, \tag{3.2}$$

where $\mu_1(m) = \sum_{k=1}^{m} \frac{k p_k}{\omega_1(m)}$ denotes the mean of the foreground and $\mu_2(m) = \sum_{k=m+1}^{L} \frac{k p_k}{\omega_2(m)}$ represents the mean of the background.

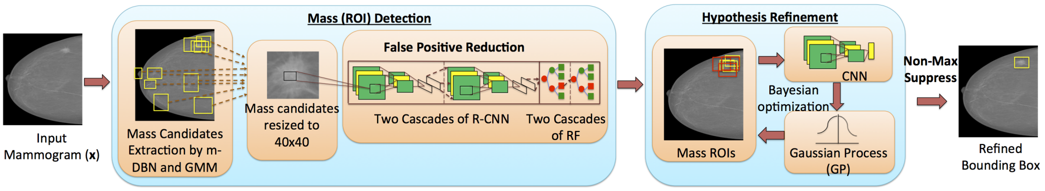


Figure 3.2: Our methodology for mass detection with refinement.

We, now define the inter-class variance σ_{A} and intra-class variance σ_{B} in terms of class variance of foreground and background:

$$\begin{aligned}\sigma_{\mathrm{A}}(m) &= \omega_1(m)\sigma_1(m) + \omega_2(m)\sigma_2(m),\\ \sigma_{\mathrm{B}}(m) &= \omega_1(m)(\mu_1(m)-\mu)^2 + \omega_2(m)(\mu_2(m)-\mu)^2\\ &= \omega_1(m)\omega_2(m)(\mu_2(m)-\mu_1(m))^2,\end{aligned} \tag{3.3}$$

where $\mu = \sum_{k=1}^{L} kp_k$ is the mean of the image $\mathbf{x}$. The optimal threshold m^* that clusters the image pixel values into foreground and background is obtained by maximising the criterion $\lambda(m)$ [129] as:

$$m^* = \arg\max_{1\geq m\geq L} \lambda(m), \tag{3.4}$$

where $\lambda(m) = \sigma_{\mathrm{B}}(m)/\sigma_{\mathrm{A}}(m)$.

In this thesis, we use the Otsu's thresholding for breast profile segmentation as shown in the Fig. 3.1.

3.4 Mass Detection

Our mass detection system [46] consists of four modules: 1) mass candidate generation, 2) false positives reduction with region based convolutional neural network (R-CNN) [63], 3) candidate selection using random forest (RF) [70], and 4) detection refinement using Bayesian optimisation [64]. The first module combines multi-scale deep belief network (m-DBN) [46, 68, 69] with Gaussian mixture model (GMM) [46, 130] to generate the mass candidate regions. These regions form the input to the second module, which consists of a two-stage cascade of R-CNNs for false positive rate reduction. The selection of the final mass candidates is performed using a two stage cascade of RF classifiers. In the fourth module, we refine mass candidates location and scale using a local search algorithm based on Bayesian optimisation [64]. In the following sections, we describe each of these modules in detail.

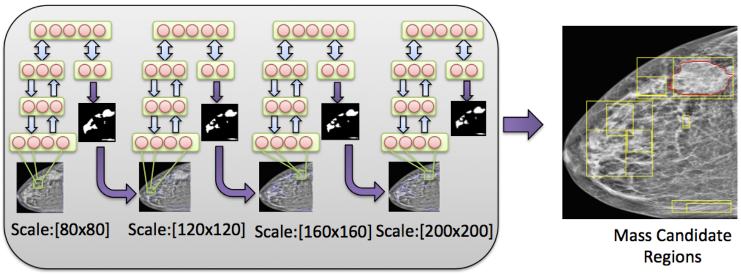


Figure 3.3: Mass candidate generation using m-DBN.

3.4.1 Mass Candidate Generation

The mass candidates generation using m-DBN [46, 68, 69, 106, 107], as shown in Fig. 3.3, consists of producing a set of N_{MDBN} bounding boxes $\{\widetilde{\mathbf{d}}_n\}_{n=1}^{N_{\text{MDBN}}}$, and coarse segmentation masks $\{\widetilde{\mathbf{y}}_n\}_{n=1}^{N_{\text{MDBN}}}$ for a mammogram $\mathbf{x}$, represented by

$$\{\widetilde{\mathbf{d}}_n, \widetilde{\mathbf{y}}_n\}_{n=1}^{N_{\text{MDBN}}} = f_{\text{MDBN}}(\mathbf{x}, \theta_{\text{MDBN}}), \tag{3.5}$$

where f_{MDBN} is the mass candidate generation model defined by the parameters θ_{MDBN}. The m-DBN model is a cascade of deep belief network (DBN) at several image resolutions, which uses a grid search on a coarse resolution of image $\mathbf{x}$ (using the mask created by Otsu's segmentation[129] as described in Sec. 3.3.1), where each grid location is classified into positive or negative based on a patch of fixed size $S \times S$ extracted around that grid location. The inference in the DBN model is based on the maximisation of the conditional probability function that is represented as:

$$P_{\text{DBN}}(\mathbf{y}(v)|\mathbf{x}_S(v), \theta_{\text{DBN}}) \propto \sum_{\mathbf{h}_1} ... \sum_{\mathbf{h}_Q} P(\mathbf{x}_S(v), \mathbf{y}(v), \mathbf{h}_1, ..., \mathbf{h}_Q), \tag{3.6}$$

where $\mathbf{x}_S(v)$ denotes the patch extracted from image $\mathbf{x}$, around the grid position v of size $S \times S$ pixels and parameter θ_{DBN} (Please note that θ_{MDBN} represents the set of parameters of several DBNs at various image scales, whereas θ_{DBN} represents the parameter of single DBN) with the DBN model consisting of a network with Q layers denoted by (below, we drop the dependence on θ_{DBN} for notation simplicity):

$$P(\mathbf{x}_S(v), \mathbf{y}(v), \mathbf{h}_1, ..., \mathbf{h}_Q) = \quad P(\mathbf{h}_Q, \mathbf{h}_{Q-1}, \mathbf{y}(v)) \left(\prod_{q=1}^{Q-2} P(\mathbf{h}_{q+1}|\mathbf{h}_q) \right) P(\mathbf{h}_1|\mathbf{x}_S(v)), \tag{3.7}$$

where $\mathbf{h}_q \in \mathbb{R}^{|q|}$ represents the hidden variables at layer q containing $|q|$ nodes. The first term in Eq. 3.7 is defined by:

$$-\log\left(P(\mathbf{h}_Q, \mathbf{h}_{Q-1}, \mathbf{y}(v))\right) \propto -\mathbf{b}_Q^\top \mathbf{h}_Q - \mathbf{a}_{Q-1}^\top \mathbf{h}_{Q-1} - \mathbf{a}_y^\top \mathbf{y}(v) - \mathbf{h}_Q^\top \mathbf{W} \mathbf{h}_{Q-1} - \mathbf{h}_Q^\top \mathbf{W}_y \mathbf{y}(v), \tag{3.8}$$

where $\mathbf{a}$, $\mathbf{b}, \mathbf{W}$ are the network parameters, and the conditional probabilities are factorized as $P(\mathbf{h}_{q+1}|\mathbf{h}_q) = \prod_{i=1}^{|q+1|} P(\mathbf{h}_{q+1}(i)|\mathbf{h}_q)$ because the nodes in layer $q+1$ are independent from each other given $\mathbf{h}_q$, which is a consequence of the DBN structure ($P(\mathbf{h}_1|\mathbf{x}_S(v))$ is similarly defined). Furthermore, each node is activated by a sigmoid function $\sigma(t) = \frac{1}{1+\exp(-t)}$, which means that $P(\mathbf{h}_{q+1}(i)|\mathbf{h}_q) = \sigma(\mathbf{b}_{q+1}(i) + \mathbf{W}_i \mathbf{h}_q)$. The learning of the DBN parameters θ_{DBN} in Eq. 3.6 is achieved with an iterative layer by layer training of auto-encoders using contrastive divergence [68, 131]. The inference is run at every position of the grid (i.e., every discrete position that falls within the breast mask of the respective image resolution) using the mean field approximation of the values in all DBN layers, which is followed by the computation of the free energy on the top layer [69]. All the points that are classified as positives on the first stage of the m-DBN are passed to the next finer resolution stage to be classified in a similar manner and this process is repeated for three coarse to fine stages, where the image resolution increases steadily at each new stage. The training of this m-DBN model at each resolution uses a training set of positive patches extracted from the grid locations that contain a pixel belonging to a mass, and negative patches from the false positive detections of previous stages, where the first coarse stage uses randomly sampled negative patches at grid locations that do not contain annotated masses. All the points that are classified as positive in each stage are combined using connected component analysis (CCA), where the similarity measure is based on the distance between the detected pixels [46].

In addition to m-DBN, we use a GMM model [130] for the generation of N_{GMM} mass candidates, represented by the bounding boxes $\{\widetilde{\mathbf{d}}_n\}_{n=1}^{N_{\text{GMM}}}$ bounding boxes and coarse segmentation masks $\{\widetilde{\mathbf{y}}_n\}_{n=1}^{N_{\text{GMM}}}$ for mammogram $\mathbf{x}$ as follows

$$\{\widetilde{\mathbf{d}}_n, \widetilde{\mathbf{y}}_n\}_{n=1}^{N_{\text{GMM}}} = f_{\text{GMM}}(\mathbf{x}, \theta_{\text{GMM}}), \tag{3.9}$$

where f_{GMM} is the GMM model defined by the parameters θ_{GMM}. This GMM model works on mammogram $\mathbf{x}$ as a pixel-wise classifier, followed by CCA. During inference, the GMM model estimates the conditional probability $P_{\text{GMM}}(\mathbf{y}(v) = 1|\mathbf{x}(v), \theta_{\text{GMM}})$ that a pixel grey value represents part of a breast mass:

$$P_{\text{GMM}}(\mathbf{y}(v) = 1|\mathbf{x}(v), \theta_{\text{GMM}}) = \frac{1}{Z} \sum_{m=1}^{M} \pi_m \mathcal{N}(\mathbf{x}(v); \mathbf{y}(v) = 1, \mu_m, \sigma_m) P(\mathbf{y}(v) = 1), \tag{3.10}$$

where $\mathcal{N}(.)$ is the normal distribution with mean value μ_m, variance Σ_m and weight π_m for each

$m = \{1, 2, .., M\}$ mixture component, Z is the normaliser that requires the computation of the conditional background probability $P_{\text{GMM}}(\mathbf{y}(v) = 0|\mathbf{x}(v), \theta_{\text{GMM}})$ and $P(\mathbf{y}(v) = 1) = 0.5$. The parameters of GMM are learned with the Expectation-Maximization (EM) algorithm [130] from the annotated training samples. The EM algorithm [130] estimates the parameter $\theta_{\text{GMM}} = [\mu_m, \Sigma_m, \pi_m]$ of the GMM model by maximising the log likelihood function using the training data. The maximum likelihood estimation for GMM using EM algorithm is carried out in two steps, namely expectation (E) step and maximization (M) step [130]. In E-step, the current values of parameters are used to determine the posterior probability of each mixture model which is subsequently used to estimate the new values of parameters in M-step. In particular, the number of components M of GMM is set at two, where we use random sampling to initialise the EM and run the algorithm for 100.

Finally, we join the candidates detected by m-DBN and GMM using the union operator to form a set of mass candidate regions $\{\widetilde{\mathbf{d}}_n, \widetilde{\mathbf{y}}_n\}_{n=1}^{N_{\text{CAN}}}$, represented by

$$\{\widetilde{\mathbf{d}}_n, \widetilde{\mathbf{y}}_n\}_{n=1}^{N_{\text{CAN}}} = \{\widetilde{\mathbf{d}}_n, \widetilde{\mathbf{y}}_n\}_{n=1}^{N_{\text{MDBN}}} \cup \{\widetilde{\mathbf{d}}_n, \widetilde{\mathbf{y}}_n\}_{n=1}^{N_{\text{GMM}}}. \tag{3.11}$$

3.4.2 False Positives Reduction with R-CNN

The set of mass candidates $\{\widetilde{\mathbf{d}}_n, \widetilde{\mathbf{y}}_n\}_{n=1}^{N_{\text{CAN}}}$ from Eq. 3.11 typically contains a large number of false positives that are then removed by the second stage of the mass detection method comprising two cascade stages of region based convolutional neural networks (R-CNN) [46, 63]. R-CNN extracts the features from the last layer of a CNN and these features are classified using linear support vector machine (SVM) [46, 63]. The false positive reduction using a cascade of R-CNN detectors is represented by

$$\{\widetilde{\mathbf{d}}_n, \widetilde{\mathbf{y}}_n\}_{n=1}^{N_{\text{RCNN}}} = f_{\text{RCNN}}(\mathbf{x}, \{\widetilde{\mathbf{d}}_n, \widetilde{\mathbf{y}}_n\}_{n=1}^{N_{\text{CAN}}}, \theta_{\text{RCNN}}), \tag{3.12}$$

where, $f_{\text{RCNN}}(.)$ denotes the function that defines the R-CNN model with parameters θ_{RCNN} (i.e., weights and biases for CNN and linear SVM), and $N_{\text{RCNN}} \leq N_{\text{CAN}}$ (i.e., the number of detections tends to reduce after this stage). A CNN [60, 61] model for a input $\widehat{\mathbf{x}}$ is a feedforward neural network and consists of multiple processing stages, represented by the following function:

$$f_{\text{CNN}}(\widehat{\mathbf{x}}, \theta_{\text{CNN}}) = f_{\text{out}} \circ g_L \circ h_L \circ f_L \circ \ldots g_1 \circ h_1 \circ f_1(\widehat{\mathbf{x}}), \tag{3.13}$$

where $f_l(.)$ represents the pre-activation stage that linearly transforms the input, θ_{CNN} is the parameter set formed by the set of K linear filters $\{\mathbf{w}_l^{(i)}\}_{i=1}^K$, and biases $\{b_l^{(i)}\}_{i=1}^K$ for each layer $l \in \{1, ..., L\}$, h_l denotes the non-linear activation function (e.g., sigmoid defined as in Sec. 3.4.1, or rectified linear unit [132]), g_l represents a non-linear sub-sampling function that pools (using either the mean or max functions) the values from a region from the input data, and

f_{out} represents the fully connected layer that is defined with a linear filter that uses all the inputs to produce one of the output values. Inference consists of the application of this process in a feed-forward manner by computing the output from Eq. 3.13, and training is carried out with stochastic gradient descent to minimise the cross entropy loss over the training set (via back propagation) [60, 61], represented by

$$\ell_{\text{CNN}}(\theta_{\text{CNN}}) = \sum_{i=1}^{|\mathcal{D}|} \sum_{n=1}^{N_{\text{CAN}}} c_{(i,n)} \log \tilde{c}_{\text{cnn}(i,n)}, \tag{3.14}$$

where θ_{CNN} is the parameter of the CNN model $\tilde{c}_{\text{cnn}(i,n)}$ is the class label predicted by this model (defined as $c = 1$ for mass and $c = 0$ for background).

The training of SVM in R-CNN in Eq. 3.12 proceeds by first cropping the mass candidate with a bounding box around the candidate region in Eq. 3.11 from the input image $\mathbf{x}$ (note that these bounding boxes are loosely localised with intersection over union ratio (IoU) =0.3 with respect to the annotated mass), resizing the cropped patch to a fixed size of $M \times M$ pixels using bi-cubic interpolation and preprocessing it with the contrast enhancement using CLAHE (described in Sec. 2.1 [47]). We use features from the final fully connected layer of the CNN model with its learned parameter θ_{CNN} in a linear SVM that is trained by minimising the following hinge loss [46, 63]

$$\ell_{\text{SVM}}(\theta_{\text{SVM}}) = \sum_{i=1}^{|\mathcal{D}|} \sum_{n=1}^{N_{\text{CAN}}} \max\left(0, 1 - c_{(i,n)}\tilde{c}_{\text{svm}(i,n)}\right), \tag{3.15}$$

where θ_{SVM} is the parameter of the linear SVM model and $\tilde{c}_{\text{svm}(i,n)}$ is the class label predicted by this model. All candidates surviving the first cascade of the R-CNN are then passed through to a second stage of R-CNN to further reduce the false positive rate as shown in Fig. 3.2. The hyper-parameters of the learning process, such as learning rate , number of filters and sizes of filters for the R-CNN are cross-validated using the validation set, where the filter weights are initialised randomly.

3.4.3 Candidate Selection using RF

The cascade of R-CNN models described in Sec. 3.4.2 still produces a large false positive rate, so we need to apply another round of false positive reduction methods. Therefore, we first extract a large number of hand-crafted features from the mass candidate regions detected by the methods described in Sec. 3.4.2, and feed them to a cascade of random forest (RF) classifiers [70], represented by

$$\{\widetilde{\mathbf{d}}_n, \widetilde{\mathbf{y}}_n\}_{n=1}^{N} = f_{\text{RF}}(\mathbf{x}, \{\widetilde{\mathbf{d}}_n, \widetilde{\mathbf{y}}_n\}_{n=1}^{N_{\text{RCNN}}}, \theta_{\text{RF}}), \tag{3.16}$$

where $f_{\text{RF}}(.)$ is the function that defines the RF classifier model with parameters θ_{RF} (i.e., number of trees, number of leaves, and feature/threshold per tree node), and $N \leq N_{\text{RCNN}}$ (i.e., the number of candidates tend to reduce after this stage). The hand-crafted features that are used at this stage can be divided into morphological features, appearance features and texture features [4, 35, 44, 48, 50, 59, 82, 93, 94]. The morphological features are based on the segmentation contour and they represent the geometrical properties of segmentation, such as margin spiculation and sharpness, area, circularity, rectangularity, perimeter, perimeter to area ratio, and normalised area length (NRL) features, such as boundary roughness, mean, entropy, standard deviation and area ratio zero crossing count. The appearance features include the ratio of grey scale values inside and in the vicinity of the segmented object. The texture features are based on spatial grey level dependence (SGLD) matrix, which is defined as a distribution of occurrence of pixel intensity i with respect to a pixel intensity j, which varies with inter-pixel separation and direction. The texture features that are based on SGLD matrix are the following: energy, correlation, inertia, entropy, difference of moment, inverse difference of moment, sum average, sum entropy, difference entropy, sum variance, difference variance, difference average, and information measure of correlation.

3.4.4 Mass Detection Refinement using Bayesian Optimisation

The final module of our mass detection system is the detection refinement step using Bayesian optimisation [64]. The primary motivation of using Bayesian optimisation in this work is that it presented superior performance compared to other local search algorithms such as hill climbing, local random search and selective search [64]. In this step, our objective is to fit the bounding boxes detected from the cascade of RF classifiers $\{\widetilde{\mathbf{d}}_n, \widetilde{\mathbf{y}}_n\}_{n=1}^{N}$ more precisely around the breast masses. Let us assume that we have some scoring function, represented by a CNN model defined as :

$$\widetilde{f}_n = f_{\text{score}}(\widetilde{\mathbf{d}}_n, \mathbf{x}, \theta_{\text{score}}), \tag{3.17}$$

which takes the bounding box $\widetilde{\mathbf{d}}_n$, mammogram $\mathbf{x}$, and parameter θ_{score} and outputs a score $\widetilde{f}_n$. We can then build a current observation set $\mathcal{B}_N = \{(\widetilde{\mathbf{d}}_n, \widetilde{f}_n)\}_{n=1}^{N}$ with the objective of finding a new bounding box $\widetilde{\mathbf{d}}_{N+1}$ that maximises score f_{N+1}, in which f is sampled from the distribution $p(f|\mathcal{B}_N) \propto p(\mathcal{B}_N|f)p(f)$. This new bounding box and score value are then used to update the hypotheses set as follows:

$$\mathcal{B}_{N+1} = \mathcal{B}_N \cup (\widetilde{\mathbf{d}}_{N+1}, f_{N+1}). \tag{3.18}$$

The objective function that we maximise under the Bayesian optimisation is called expected improvement (a_{ei}) and is defined as [64]:

$$a_{\text{ei}}(\widetilde{\mathbf{d}}_{N+1}|\mathcal{B}_N,\theta_{\text{GP}}) = \int_{\hat{f}_N}^{\infty} (f_{N+1} - \hat{f}_N).P_{\text{GP}}(f_{N+1}|\widetilde{\mathbf{d}}_{N+1},\mathcal{B}_N;\theta_{\text{GP}})df, \tag{3.19}$$

where $P_{\text{GP}}(f_{N+1}|\widetilde{\mathbf{d}}_{N+1},\mathcal{B}_N;\theta_{\text{GP}})$ is a Gaussian Process (GP) [64, 128] prior with parameter θ_{GP}, and $\hat{f}_N = \max_{n\in\{1,..,N\}} f_n$. The parameter θ_{GP} in Eq. 3.19 denotes the mean kernel $m : \mathbb{R}^4 \rightarrow \mathbb{R}$, a positive definite covariance kernel $k : \mathbb{R}^4 \times \mathbb{R}^4 \rightarrow \mathbb{R}$ and small Gaussian noise β that is added to f_n for numerical stability. The value of mean kernel m_0 in GP is fixed, but the covariance kernel is updated as follows [64, 128]:

$$k(\widetilde{\mathbf{d}}(v),\widetilde{\mathbf{d}}(q);z) = \eta \exp\,(\frac{1}{2}(\Phi(\widetilde{\mathbf{d}}(v)) - \Phi(\widetilde{\mathbf{d}}(q)))^T\Lambda(\Phi(\widetilde{\mathbf{d}}(v)) - \Phi(\widetilde{\mathbf{d}}(q)))), \tag{3.20}$$

where Λ is a 4×4 diagonal matrix whose diagonal entries, $\lambda^2_{i=1,..,4}$ along with m, η forms a seven dimensional hyper-parameter of GP ($\theta_{\text{GP}} = [\beta, m_0, \eta, \lambda_1^2, \lambda_2^2, \lambda_3^2, \lambda_4^2]$) that is learned from the training data, $\widetilde{\mathbf{d}}(v), \widetilde{\mathbf{d}}(q)$ are the v, q elements of bounding box $\widetilde{\mathbf{d}}$, and $\Phi : \mathbb{R}^4 \rightarrow \mathbb{R}^4$ parametrises the bounding box coordinates $\widetilde{\mathbf{d}}$ into a form given by [64]:

$$\Phi(\widetilde{\mathbf{d}}) = \left[\frac{x+w}{2\exp{(z)}}; \frac{y+h}{2\exp{(z)}}; \log\, w; \log\, h\right], \tag{3.21}$$

where $\frac{x+h}{2}$, $\frac{y+w}{2}$ denotes the centre coordinates of the given bounding box, w, h are the width, height of the bounding box, and z is the latent variable that has been introduced to make the covariance kernel scale invariant [64]. We obtain the hyper-parameter z in a data-driven way by maximising the marginal likelihood P_{ML} of $\mathcal{B}_N$ as [64]:

$$\tilde{z} = \arg\max_z P_{\text{ML}}(\{\widetilde{f}\}_{i=1}^N|\{\widetilde{\mathbf{d}}\}_{i=1}^N;\theta_{\text{GP}}). \tag{3.22}$$

The estimation of the parameter θ_{score} for the scoring function in Eq. 3.17 is done by training a CNN model using manually annotated bounding boxes $\mathbf{d}$ from training dataset $\mathcal{D}$. The ground truth bounding boxes are artificially augmented (with random translation and scale) such that these artificial positive bounding boxes have intersection over union (IoU) ratio above a pre-defined threshold ρ with respect to the manual annotation, and the negative samples have IoU below that same threshold. The bounding boxes are cropped, resized to $M \times M$ and pre-processed with contrast enhancement technique [47] (described in Chapter 2, Sec 2.1). Similarly, the estimation of the parameter θ_{GP} of the GP model is done using ground truth bounding boxes $\mathbf{d}$ and their scores f using Eq. 3.17 from the training set $\mathcal{D}$ by maximising the joint likelihood of such observations [128] as:

Algorithm 3.1: Local Search for Detection Refinement

```
Input: Mammogram $\mathbf{x}$, the set of detected bounding boxes, and scores $\mathcal{B}_N = \{(\widetilde{\mathbf{d}}_n, \widetilde{f}_n)\}_{n=1}^N$,
    parameters $\theta_{\text{score}}$ for the scoring function in Eq. 3.17, acquisition function parameters $\theta_{\text{GP}}$
    in Eq. 3.19, maximum number of iterations $t_{\text{max}}$, and the threshold $f_{\text{prune}}$ to prune the
    bounding boxes.
$\mathcal{B}_{\text{new}} \leftarrow \text{transformations}(\mathcal{B}_N)$
for $t = 1, ..., t_{\text{max}}$ do
    $\mathcal{B}_{\text{proposal}} = \emptyset$
    $\mathcal{B}_{\text{prune}} = \{(\mathbf{d}, f) \in \mathcal{B}_{\text{new}} : f \geq f_{\text{prune}}\}$
    $\mathcal{B}_{\text{nms}} = \text{NMS}(\mathcal{B}_{\text{prune}})$
    for $(\mathbf{d}_{\text{best}}, f_{\text{best}}) \in \mathcal{B}_{\text{nms}}$ do
        for $\rho \in \{0.3, 0.5, 0.7\}$ do
            $\mathcal{B}_{\text{local}} = \{(\mathbf{d}, f) \in \mathcal{B}_{\text{new}} : \text{IoU}(\mathbf{d}, \mathbf{d}_{\text{best}}) > \rho\}$
            $\tilde{z} = \arg\max_z P_{\text{ML}}(\mathcal{B}_{\text{local}}; \theta_{\text{GP}})$
            $\mathbf{d}_{N+1} = \arg\max_{\mathbf{d}} a_{\text{ei}}(\mathbf{d}|\mathcal{B}_{\text{local}}, \theta_{\text{GP}}, \tilde{z})$
            $f_{N+1} = f_{\text{score}}(\mathbf{d}_{N+1}, \mathbf{x}; \theta_{\text{score}})$
            $\mathcal{B}_{\text{proposal}} \leftarrow \mathcal{B}_{\text{proposal}} \cup (\mathbf{d}_{N+1}, f_{N+1})$
        end for
    end for
    $\mathcal{B}_{\text{new}} \leftarrow \mathcal{B}_{\text{proposal}} \cup \mathcal{B}_{\text{new}}$
end for
$\mathcal{B}_{\text{prune}} = \{(\mathbf{d}, f) \in \mathcal{B}_{\text{new}} : f \geq f_{\text{prune}}\}$
$\mathcal{B}_{\text{ref}} = \text{NMS}(\mathcal{B}_{\text{prune}})$
```

$$\theta_{\text{GP}}^* = \arg\max_{\theta_{\text{GP}}} \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{A}|_i} \log \, P_{\text{GP}}(\mathbf{d}_{(i,j)}, f_{(i,j)}; \theta_{\text{GP}}). \tag{3.23}$$

The detection refinement runs according to the steps in Algorithm 3.1, where the transformations(.) function translates and scales the samples in $\mathcal{B}_N$ to form the set $\mathcal{B}_{\text{new}}$ and non-max suppression (NMS) is a function that takes a set of bounding boxes with their scores and classify them by suppressing all the bounding boxes with similar IoU ratio based on their scores (i.e., bounding boxes with the low scores are removed). The detection refinement in Algorithm 3.1 continues for fixed number of iterations t_{max}, the algorithm prunes candidates with low scores using the threshold f_{prune}, and cluster the remaining candidates using NMS. For each bounding box $(\mathbf{d}_{\text{best}}, f_{\text{best}})$ that has been considered to be a local optimum using NMS, the algorithm considers the range of IoU values ($\rho \in \{0.3, 0.5, 0.7\}$) to build the local bounding box set $\mathcal{B}_{\text{local}}$. The newly formed local observation set $\mathcal{B}_{\text{local}}$ is used in the GP to sample the new bounding box $\mathbf{d}_{N+1}$, which integrates the new set of proposals. This process outputs the set $\mathcal{B}_{\text{ref}}$ of final mass candidates. In practice, we set $t_{max} = 10$, resulting in an average running time of 10s for the local search Algorithm 3.1.

In Chapter 9, we present the results of bounding box refinement based on the local search in Algorithm 3.1 using Bayesian optimisation.

3.5 Mass Segmentation

In this section, we describe the structured output prediction models for breast mass segmentation, which is one of the main contribution of this thesis. We start this section with an explanation of the learning process in our two structured output prediction models [133], namely: 1) conditional random field (CRF) which uses truncated fitting [72] for learning the model parameters and tree re-weighted belief propagation (TRW) [72, 73] for inference, and 2) structured support vector machines (SSVM) [74, 75] which uses SSVM for learning the model parameters and graph cuts [75, 80] for inference (see Fig. 3.4 [39]). The segmentation using these structured output models is obtained in low resolution sub-image $\widetilde{\mathbf{x}}$ containing a mass candidate. In the automated set-up, we use each bounding box $\mathbf{d}_n \in \mathcal{B}_{\text{ref}}$ estimated from the hypothesis refinement in Alg. 3.1 (Chapter 9), whereas in the manual set-up, we obtain $\mathbf{d}_n$ by extracting a bounding box from around the centre of the manual annotation of the test/train image, where the size for each dimension of the rectangle is obtained using the size of the annotation plus two pixels [39, 106, 107] (Chapters 4, 5, and 6). The image patch extracted from this bounding box is then resized to an image of size $M \times M$ pixels with the function $\widetilde{\mathbf{x}}_n = f_{\text{crop}}(\mathbf{x}, \mathbf{d}_n)$ using bi-cubic interpolation and pre-processed using the contrast enhancement technique described in Sec. 2.1 [47]. Both of these structured output prediction models are represented in terms of graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the nodes and $\mathcal{E}$ represents the edges of the graph. The learning of the parameter θ_{SP} of the structured prediction models is carried out by minimising a continuous convex loss function ℓ as [133]:

$$\theta^*_{\text{SP}} = \arg\min_{\theta_{\text{SP}}} \sum_{i=1}^{|\mathcal{D}|} \sum_{n=1}^{|\mathcal{B}_{\text{ref}}(i)|} \ell(\widetilde{\mathbf{x}}_{i,n}, \widetilde{\mathbf{y}}_{i,n}, \theta_{\text{SP}}), \tag{3.24}$$

where i indexes the training images from set $\mathcal{D}$ and n indexes the masses in the set of refined detections $\mathcal{B}_{\text{ref}}$ (with cardinality $|\mathcal{B}_{\text{ref}}|$), $\widetilde{\mathbf{y}}_{n,i}$ denotes the cropped segmentation map obtained with $f_{\text{crop}}(\mathbf{y}_i, \mathbf{d}_n)$, defined above, and $\ell(\widetilde{\mathbf{x}}_{i,n}, \widetilde{\mathbf{y}}_{i,n}, \theta_{\text{SP}})$ is a continuous and convex loss function, defined below. The optimisation problem in Eq. 3.24 can be solved in many different ways, but in this thesis we make use of two methods, described in Sec. 3.5.1 and Sec. 3.5.2. Hereafter, we drop the indices n for notation simplicity.

3.5.1 Conditional Random Field (CRF)

In the CRF setup , the loss function in Eq. 3.24 is defined with the energy term $E(\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}; \theta_{\text{SP}})$ and log-partition function $A(\widetilde{\mathbf{x}}; \theta_{\text{SP}})$:

$$\ell(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}, \theta_{\text{SP}}) = A(\widetilde{\mathbf{x}}, \theta_{\text{SP}}) - E(\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}; \theta_{\text{SP}}), \tag{3.25}$$

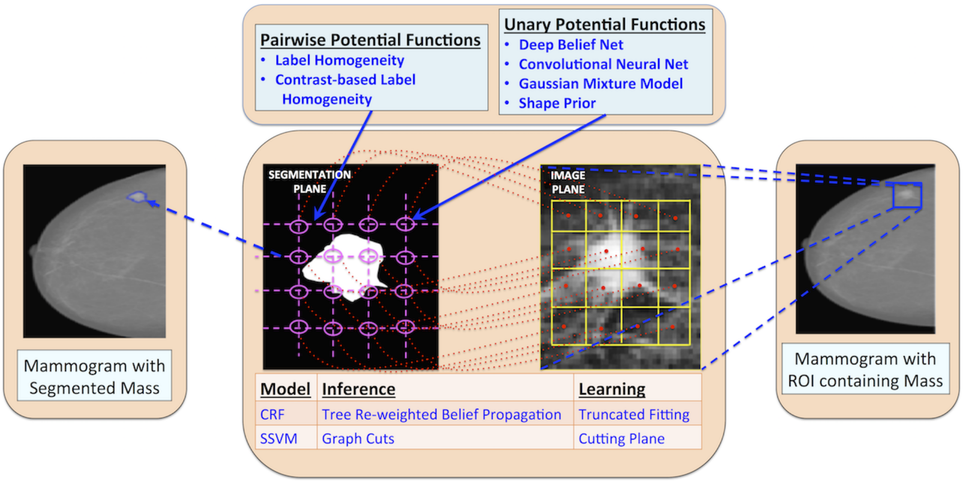

Pairwise Potential Functions
• Label Homogeneity
• Contrast-based Label Homogeneity
Unary Potential Functions
• Deep Belief Net
• Convolutional Neural Net
• Gaussian Mixture Model
• Shape Prior
SEGMENTATION PLANE
IMAGE PLANE
Mammogram with Segmented Mass
Mammogram with ROI containing Mass
Model
Inference
Learning
CRF
Tree Re-weighted Belief Propagation
Truncated Fitting
SSVM
Graph Cuts
Cutting Plane

Figure 3.4: Structured Prediction Model for segmentation of masses.

where $A(\tilde{\mathbf{x}};\theta_{\text{SP}}) = \log \sum_{\tilde{\mathbf{y}}\in\{-1,+1\}^{M\times M}} \exp\,\{E(\tilde{\mathbf{y}},\tilde{\mathbf{x}};\theta_{\text{SP}})\}$ is the normalizer, and

$$E(\tilde{\mathbf{x}},\tilde{\mathbf{y}};\theta_{\text{SP}}) = \sum_{k=1}^{K}\sum_{v\in\mathcal{V}} \theta_{1,k}\psi^{(1,k)}(\tilde{\mathbf{y}}(v),\tilde{\mathbf{x}}) + \sum_{l=1}^{L}\sum_{v,q\in\mathcal{E}} \theta_{2,l}\psi^{(2,l)}(\tilde{\mathbf{y}}(v),\tilde{\mathbf{y}}(q),\tilde{\mathbf{x}}), \tag{3.26}$$

where $\psi^{(1,k)}(.,.)$ denotes one of the K potential functions between label and pixel nodes (please refer to Fig. 3.4), $\psi^{(2,l)}(.,.,.)$ represents one of the L potential functions on the edges between label nodes, $\theta_{\text{SP}} = [\theta_{1,1},...,\theta_{1,K},\theta_{2,1},...,\theta_{2,L}]^{\top} \in \mathbf{R}^{K+L}$, and $\tilde{\mathbf{y}}(v),\tilde{\mathbf{y}}(q)$ are the v^{th} and q^{th} components of vector $\tilde{\mathbf{y}}$. We minimise the loss function in Eq. 3.24 using tree re-weighted belief propagation, which provides the upper bound to the log partition function $A(\tilde{\mathbf{x}};\theta_{\text{SP}})$ in Eq. 3.25 [73]:

$$A(\tilde{\mathbf{x}};\theta_{\text{SP}}) = \max_{\mu\in\mathcal{M}} \theta_{\text{SP}}^{T}\mu + H(\mu), \tag{3.27}$$

where $\mathcal{M} = \{\mu' : \exists\theta_{\text{SP}}, \mu' = \mu\}$ is the marginal polytope, $\mu = \sum_{\tilde{\mathbf{y}}\in\mathbf{m}\in\{-1,+1\}^{M\times M}} P(\tilde{\mathbf{y}}|\tilde{\mathbf{x}},\theta_{\text{SP}})$ $f(\tilde{\mathbf{y}})$, $f(\tilde{\mathbf{y}})$ is the set of indicator functions of possible configurations of each clique and variable in the graph [134] (as in Eq. 3.26), $P(\tilde{\mathbf{y}}|\tilde{\mathbf{x}},\theta_{\text{SP}}) = \exp\{E(\tilde{\mathbf{y}},\tilde{\mathbf{x}};\theta_{\text{SP}}) - A(\tilde{\mathbf{x}};\theta_{\text{SP}})\}$ indicates the conditional probability of the annotation $\tilde{\mathbf{y}}$ given the image $\tilde{\mathbf{x}}$ and parameters θ_{SP} (assuming $P(\tilde{\mathbf{y}}|\tilde{\mathbf{x}};\theta_{\text{SP}})$ belongs to the exponential family), and entropy $H(\mu)$ is given by:

$$H(\mu) = -\sum_{\tilde{\mathbf{y}}\in\mathbf{m}\in\{-1,+1\}^{M\times M}} P(\tilde{\mathbf{y}}|\tilde{\mathbf{x}};\theta_{\text{SP}})\log\,P(\tilde{\mathbf{y}}|\tilde{\mathbf{x}},\theta_{\text{SP}}). \tag{3.28}$$

The marginal polytope $\mathcal{M}$ is difficult to define and the entropy $\mathbf{H}(\mu)$ is not tractable [72] for the general cyclic graph, which we use for the breast mass segmentation in this thesis. We solve

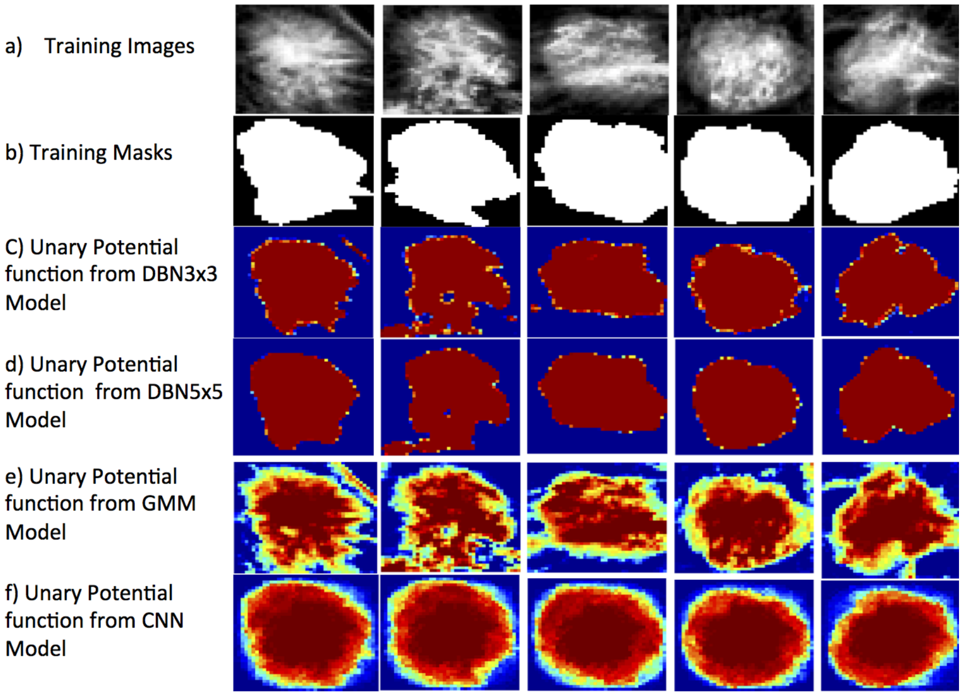

a) Training Images
b) Training Masks
C) Unary Potential function from DBN3x3 Model
d) Unary Potential function from DBN5x5 Model
e) Unary Potential function from GMM Model
f) Unary Potential function from CNN Model

Figure 3.5: Examples of unary potentials for few training images from DBN, CNN and GMM .

these problems using tree re-weighted belief propagation (TRW), which replaces the marginal polytope with a superset $\mathcal{L} \supset \mathcal{M}$ that only considers the local constraints of the marginals, and then approximates the entropy calculation with an upper bound. The learning process involves the estimation of θ_{SP} by gradient descent that minimises the loss in Eq. 3.25, which is defined by the change rate of θ_{SP} between successive gradient descent iterations. However, as noted by Domke [72], there are problems with this approach, where large thresholds in this change rate can lead to suboptimal estimations, and tight thresholds result in slow convergence. These issues are circumvented by the truncated fitting algorithm [72], which uses a fixed number of iterations (i.e., no threshold is used in this training algorithm).

3.5.2 Structured Support Vector Machine (SSVM)

In the SSVM setup, we use the following loss function:

$$\ell(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{y}}_i, \theta_{\text{SP}}) = \max_{\widetilde{\mathbf{y}} \in \mathcal{Y}} \left(\Delta(\widetilde{\mathbf{y}}_i, \widetilde{\mathbf{y}}) + E(\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}_i; \theta_{\text{SP}}) - E(\widetilde{\mathbf{y}}_i, \widetilde{\mathbf{x}}_i; \theta_{\text{SP}}) \right), \tag{3.29}$$

where $\Delta(\widetilde{\mathbf{y}}_i, \widetilde{\mathbf{y}})$ represents the dissimilarity between $\widetilde{\mathbf{y}}_i$ and $\widetilde{\mathbf{y}}$, which satistfies the conditions $\Delta(\widetilde{\mathbf{y}}_i, \widetilde{\mathbf{y}}) \geq 0$ for $\widetilde{\mathbf{y}}_i \neq \widetilde{\mathbf{y}}$ and $\Delta(\widetilde{\mathbf{y}}_i, \widetilde{\mathbf{y}}_i) = 0$.

The estimation of θ_{SP} using SSVM optimisation proceeds by formulating a regularised loss minimisation problem, which can be represented by: $\theta^*_{\text{SP}} = \min_{\theta_{\text{SP}}} \|\theta_{\text{SP}}\|^2 + \lambda \sum_i \ell(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{y}}_i, \theta_{\text{SP}})$, with $\ell(.)$ defined in Eq. 3.29. The introduction of slack variables leads to the following optimization problem [74, 75]:

$$\begin{array}{ll} \text{minimize}_{\theta_{\text{SP}}} & \frac{1}{2}\|\theta_{\text{SP}}\|^2 + \frac{C}{|\mathcal{D}|}\sum_i \xi_i \\ \text{subject to} & E(\widetilde{\mathbf{y}}_i, \widetilde{\mathbf{x}}_i; \theta_{\text{SP}}) - E(\hat{\mathbf{y}}_i, \widetilde{\mathbf{x}}_i; \theta_{\text{SP}}) \geq \Delta(\widetilde{\mathbf{y}}_i, \hat{\mathbf{y}}_i) - \xi_i, \forall \hat{\mathbf{y}}_i \neq \widetilde{\mathbf{y}}_i \\ & \xi_i \geq 0. \end{array} \tag{3.30}$$

This optimisation is a quadratic programming problem involving an intractably large number of constraints. In order to keep the number of constraints manageable, we use the cutting plane method that keeps a relatively small subset of the constraints by solving the maximisation problem:

$$\hat{\widetilde{\mathbf{y}}}_i = \arg\max_{\widetilde{\mathbf{y}}} \Delta(\widetilde{\mathbf{y}}_i, \widetilde{\mathbf{y}}) + E(\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}_i; \theta_{\text{SP}}) - E(\widetilde{\mathbf{y}}_i, \widetilde{\mathbf{x}}_i; \theta_{\text{SP}}) - \xi_i, \tag{3.31}$$

which finds the most violated constraint for the i^{th} training sample given the parameter θ_{SP}. Then if the right hand side is strictly larger than zero, the most violated constraint is included in the constraint set and Eq. 3.30 is re-solved. This iterative process runs until no more constraints are found. Note that if we remove the constants from Eq. 3.31, the optimization problem is simply: $\hat{\widetilde{\mathbf{y}}}_i = \arg\max_{\widetilde{\mathbf{y}}} \Delta(\widetilde{\mathbf{y}}_i, \widetilde{\mathbf{y}}) + E(\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}_i; \theta_{\text{SP}})$, which can be efficiently solved using graph cuts [80], if the function $\Delta(.,.)$ can be properly decomposed in the label space. A simple example that works with graph cuts is $\Delta(\widetilde{\mathbf{y}}, \widetilde{\mathbf{y}}_i) = \sum_i 1 - \delta(\widetilde{\mathbf{y}}(v) - \widetilde{\mathbf{y}}_i(v))$, which represents the Hamming distance and can be decomposed in the label space. Therefore, we use it in our methodology.

The label inference for a test mammogram $\mathbf{x}$, given the learned parameters θ_{SP} from Eq. 3.30, is based on the following inference:

$$\widetilde{\mathbf{y}}^* = \arg\max_{\widetilde{\mathbf{y}}} E(\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}; \theta_{\text{SP}}), \tag{3.32}$$

which can be efficiently and optimally solved for binary problems with graph cuts [80].

3.5.3 Potential Functions

In this section we define the unary $\psi^{(1,k)}$ and pairwise potential functions $\psi^{(2,l)}$ to be used in Eq. 3.25. The unary potential function $\psi^{(1,1)}$ is represented by a shape prior [39, 106, 107], which is estimated from the training set at each image lattice position v, as follows:

$$\psi^{(1,1)}(\widetilde{\mathbf{y}}(v), \widetilde{\mathbf{x}}) = -\log\, P_{\text{prior}}(\widetilde{\mathbf{y}}(v) = 1|\theta_{\text{prior}}), \tag{3.33}$$

where $P_{\text{prior}}(\widetilde{\mathbf{y}}(v) = 1|\theta_{\text{prior}}) = 1/|\mathcal{D}|\sum_i \delta(\widetilde{\mathbf{y}}_i(v) - 1)$ and $\delta(.)$ represents the Dirac delta function.

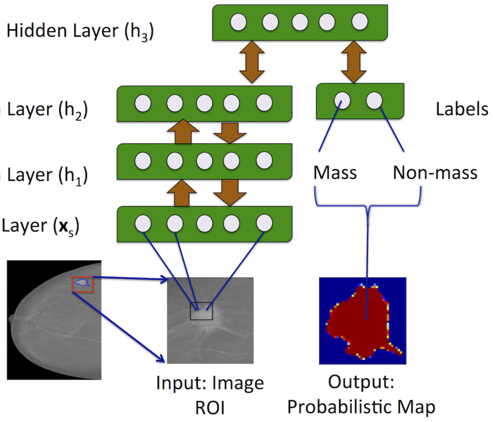


Figure 3.6: DBN model that takes the mass candidate as input and outputs a unary potential (probability map) for our segmentation algorithm (Chapters 4, 5, 6).

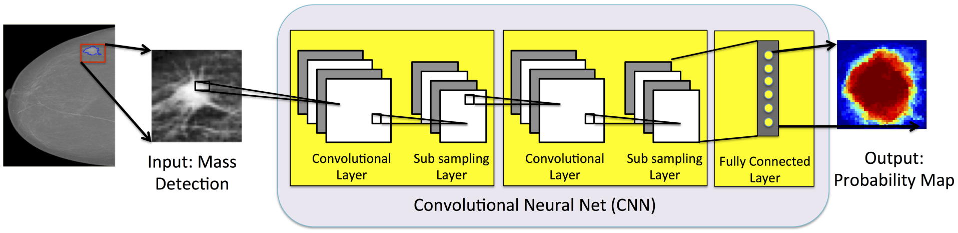


Figure 3.7: The CNN model that takes mass candidate as input and outputs a unary potential (probability map) for our segmentation algorithm (Chapter 6).

The unary potential function $\psi^{(1,2)}$ in Eq. 3.26 is based on GMM [39, 106, 107, 130] shape model represented as:

$$\psi^{(1,2)}(\tilde{\mathbf{y}}(v), \tilde{\mathbf{x}}) = -\log\, P_{\text{GS}}(\tilde{\mathbf{y}}(v) = 1|\tilde{\mathbf{x}}(v), \theta_{\text{GS}}), \tag{3.34}$$

where θ_{GS} is the parameter of GMM model, $P_{\text{GS}}(\tilde{\mathbf{y}}(v) = 1|\tilde{\mathbf{x}}(v), \theta_{\text{GS}})$ is the conditional probability distribution. The training and inference processes using a GMM model have been described using Eq. 3.10 in Sec. 3.4.1.

The potential function $\psi^{(1,3)}$ in Eq. 3.26, based on DBN [39, 69, 106, 107] shape model, is represented as:

$$\psi^{(1,3)}(\tilde{\mathbf{y}}(v), \tilde{\mathbf{x}}) = -\log\, P_{\text{DS}}(\tilde{\mathbf{y}}(v)|\tilde{\mathbf{x}}_S(v), \theta_{\text{DS}}), \tag{3.35}$$

where $\tilde{\mathbf{x}}_S(v)$ denotes the patch extracted from the sub-image $\tilde{\mathbf{x}}$, around the lattice position v of size $S \times S$ pixels, θ_{DS} is the parameter of the DBN model, $P_{\text{DS}}(\tilde{\mathbf{y}}(v)|\tilde{\mathbf{x}}_S(v), \theta_{\text{DS}})$ is the conditional probability distribution. The training and inference processes using the DBN model have been described using Eq. 3.6 in Sec. 3.4.1. In the experiment, we tried the patches of different sizes (3 × 3, 5 × 5 and 7 × 7).

The potential function $\psi^{(1,4)}$ in Eq. 3.26, based on CNN [39, 61] shape model (as shown in the Fig. 3.7), is represented by:

$$\psi^{(1,4)}(\widetilde{\mathbf{y}}(v), \widetilde{\mathbf{x}}) = -\log\, P_{\mathrm{CS}}(\widetilde{\mathbf{y}}(v) = 1|\widetilde{\mathbf{x}}, \theta_{\mathrm{CS}}), \tag{3.36}$$

where, $P_{\mathrm{CS}}(\widetilde{\mathbf{y}}(v) = 1|\widetilde{\mathbf{x}}, \theta_{\mathrm{CS}})$ represents the probability of labelling pixel v as mass or background and θ_{CS} represents the parameter of CNN model. The training and inference processes using the CNN model have been described using Eq. 3.13 in Sec. 3.4.1.

Generally, the last layer of CNN is modified to fit the particular problem of segmentation, classification or regression. For the problem of breast mass segmentation, we use the CNN (as depicted in Fig. 3.7), which has the number of nodes in last layer equal to the number of pixels in the input image and we minimise the binary segmentation error using a pixel-wise cross entropy loss, defined as

$$\ell_{\mathrm{CS}}(\theta_{\mathrm{CS}}) = \sum_{i=1}^{|\mathcal{D}|} \sum_{j}^{|\mathcal{A}_i|} \sum_{v \in M \times M} \widetilde{\mathbf{y}}_{(i,j)}(v) \log\, \widetilde{\mathbf{y}}^{*}_{(i,j)}(v), \tag{3.37}$$

where $\widetilde{\mathbf{y}}^{*}_{(i,j)}(v)$ is the pixel-wise label predicted by this model. Fig. 3.5 shows some examples of the results from the various unary potential functions (i.e., DBN, GMM and CNN) that we use with our structured output prediction models.

The pairwise binary functions between label nodes in Eq. 3.26 represent label and contrast related labelling homogeneities: $\psi^{(2,1)}(\widetilde{\mathbf{y}}(v), \widetilde{\mathbf{y}}(q), \widetilde{\mathbf{x}})$ and $\psi^{(2,1+n)}$ $(\widetilde{\mathbf{y}}(v), \widetilde{\mathbf{y}}(q), \widetilde{\mathbf{x}})$, respectively [72, 107, 133]. We define the labelling homogeneity as:

$$\psi^{(2,1)}(\widetilde{\mathbf{y}}(v), \widetilde{\mathbf{y}}(q), \widetilde{\mathbf{x}}) = 1 - \delta(\widetilde{\mathbf{y}}(v) - \widetilde{\mathbf{y}}(q)), \tag{3.38}$$

where $\delta(.)$ represents the Dirac delta function. In addition to the labelling homogeneity, we define contrast dependent labelling homogeneity by:

$$\begin{aligned} \psi^{(2,1+n)}(\widetilde{\mathbf{y}}(v), \widetilde{\mathbf{y}}(q), \widetilde{\mathbf{x}}) &= (1 - \delta(\widetilde{\mathbf{y}}(v) - \widetilde{\mathbf{y}}(q))\delta(|| \lfloor \widetilde{\mathbf{x}}(v) \rfloor_{\tau_n} - \lfloor \widetilde{\mathbf{x}}(q) \rfloor_{\tau_n} ||_2)), \\ \lfloor \widetilde{\mathbf{x}}(v) \rfloor_{\tau_n} &= \begin{cases} \widetilde{\mathbf{x}}(v) \text{ if } \widetilde{\mathbf{x}}(v) \geq \tau_n \\ 0, \text{ otherwise}, \end{cases} \end{aligned} \tag{3.39}$$

where $\widetilde{\mathbf{x}}(v), \widetilde{\mathbf{x}}(q)$ represents the grey value of the pixel at location v, q in image grid, and $\tau_n \in \{\tau_1, \tau_2, ..., \tau_{10}\}$ is a set of ten thresholds [72, 107].

In Chapters 4, 5, and 6, we compare the performance of the CRF and SSVM models with the combination of the unary and pairwise potential functions described in this chapter and report the results in terms of Dice index and running time.

```
Algorithm 3.2: Mass Segmentation with Active Contour Refinement
Input: Mammogram $\mathbf{x}$, refined bounding box $\mathbf{d}_n \in \mathcal{B}_N$, sub-image size $M_{\text{sub}}$, number of
    iterations $t_{\max}$ for the Chan-Vese optimisation, the unary and pairwise model parameters
    $\theta_{\text{CNN}}, \theta_{\text{DBN}}, \theta_{\text{GMM}}, \theta_{\text{prior}}$, and structured output model $\theta_{\text{CRF}}$
1: Obtain sub-image $\widetilde{\mathbf{x}} = f_{\text{crop}}(\mathbf{d}_n, \mathbf{x}, M \times M)$
2: Contrast enhance sub-image $\widetilde{\mathbf{x}}$ ([39, 47])
3: Compute unary potential function results $\psi^{(1,k)}$ for $k \in \{1, ..., 4\}$ using Eq. 3.33-3.35
4: Compute pairwise potentials $\psi^{(2,l)}$ for $k \in \{1, 2\}$ using ([134])
5: Infer segmentation label $\widetilde{\mathbf{y}}^*$ using TRW ([39, 73]) or graph cuts [39, 80]
6: Restore $\widetilde{\mathbf{y}}^*$ to $\widehat{\mathbf{y}}^* = f_{\text{restore}}(\widetilde{\mathbf{y}}^*, \mathbf{d}_n)$
7: Compute initial distance function $\phi_0 = f_\phi(\widehat{\mathbf{y}}^*)$
8: Estimate $\phi_{t_{\max}}$ using active contour ([29])
9: Infer final segmentation $\mathbf{y}_n^* = \phi_{t_{\max}} \geq 0$
```

3.5.4 Mass Segmentation Refinement using Active Contour Model

The main issue of segmentation $\widetilde{\mathbf{y}}^*$ using our structured output prediction models described in Sec. 3.5.1, and Sec. 3.5.2 is the fact that they are performed in a low resolution sub-image of size $M \times M$. If we restore the segmentation $\widetilde{\mathbf{y}}^*$ to the original image resolution, using the bounding box $d_n \in \mathbf{B}_{\text{ref}}$ with a function $\widehat{\mathbf{y}}^* = f_{\text{restore}}(\widetilde{\mathbf{y}}^*, \mathbf{d}_n)$ that uses nearest neighbour interpolation, then the segmentation $\widehat{\mathbf{y}}^*$ would result in a coarse edge boundary, which needs refinement (please see the pink contour in Fig. 3.8). We solve this problem with the use of Chan-Vese active contour model [29], which is initialised using the coarse segmentation $\widehat{\mathbf{y}}^*$. The active contour model is represented by a level set function $\phi(.)$ using the signed distance function and we use $\widehat{\mathbf{y}}_n^*$ to initialise the level set function $\phi_0 = f_\phi(\widehat{\mathbf{y}}^*)$, and minimise the following the energy functional ([29]):

$$E_{\text{AC}}(\phi, \widehat{\mathbf{y}}^*, \mathbf{x}) = \gamma \int_\Omega |(\mathbf{x} - c_2)|^2 (1 - H(\phi) dx + \lambda \int_\Omega |(\mathbf{x} - c_1)|^2 H(\phi) dx + \mu \int_\Omega \delta(\phi) | \bigtriangledown \phi| dx, \tag{3.40}$$

where μ, λ, γ are the hyper-parameters, c_1, c_2 are the average of the image $\mathbf{x}$ in the regions where $\phi(.) \geq 0$ and $\phi(.) < 0$ (respectively), $\delta(.)$ is the Dirac delta function, and $H(.)$ is the heavyside step function, defined as:

$$H(\phi) = \begin{cases} 1 & \phi \geq 0 \\ 0 & \text{Otherwise.} \end{cases} \tag{3.41}$$

We minimise the energy in Eq. 3.40 by finding the steady state solution of the gradient flow equation $\frac{\partial \phi}{\partial t} = -\frac{\partial E_{\text{AC}}}{\partial \phi}$, where $\frac{\partial E_{\text{AC}}}{\partial \phi}$ is the Gâteaux derivative of the functional $E_{\text{AC}}(.)$ ([29]). The final segmentation $\mathbf{y}_n^*$ is defined by the binary map from the positive region of the level set function, i.e., $\phi \geq 0$, produced by active contour refinement (notice the green contour in Fig. 3.8). The fully automated breast mass segmentation algorithm is shown in Algorithm. 3.2.

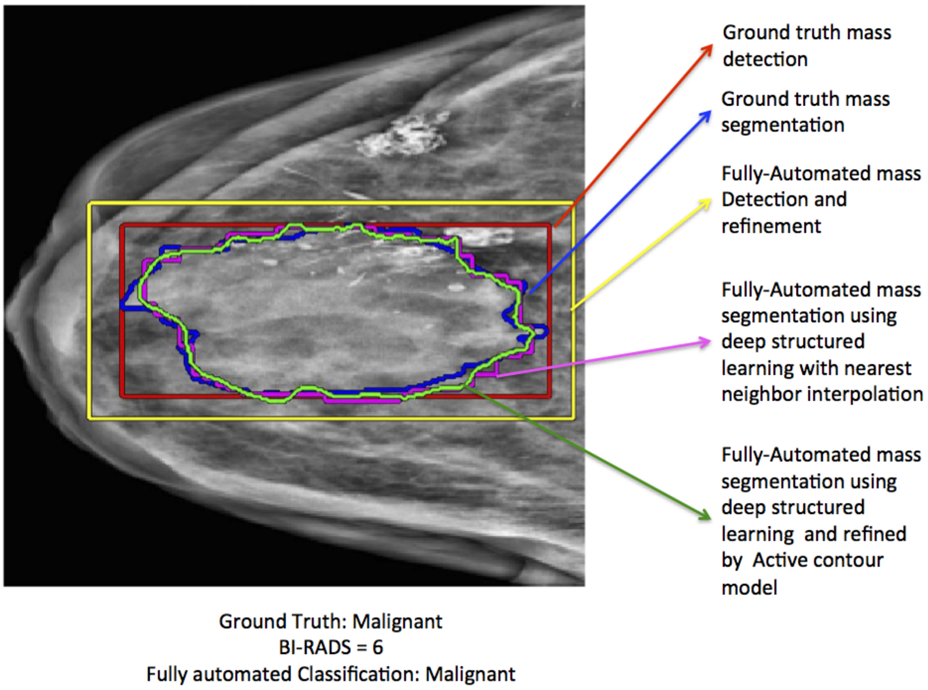

Ground truth mass detection
Ground truth mass segmentation
Fully-Automated mass Detection and refinement
Fully-Automated mass segmentation using deep structured learning with nearest neighbor interpolation
Fully-Automated mass segmentation using deep structured learning and refined by Active contour model
Ground Truth: Malignant
BI-RADS = 6
Fully automated Classification: Malignant

Figure 3.8: Example of fully automated mass detection, segmentation and classification of mass from mammograms using the proposed methodology.

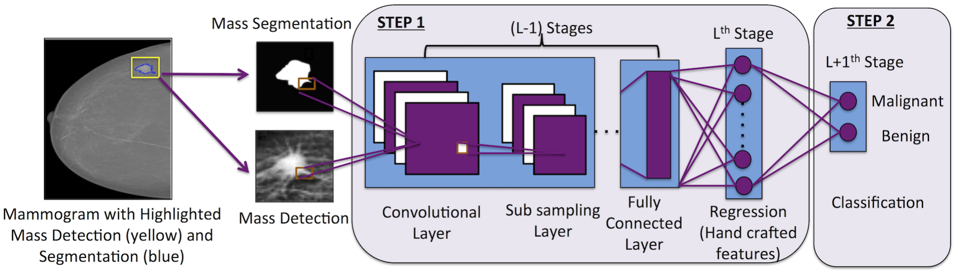


Figure 3.9: Two stage CNN model for mass classification in mammograms (Chapter 8).

In Chapter 9, we show the results of fully automated segmentation with active contour refinement and compare these results with the state-of-the-art methods in the field.

3.6 Mass Classification

The final stage of our fully automated system for the analysis of breast masses is the classification of such masses into malignant/benign. Our classification is based on a transfer learning approach using deep learning [61, 81], consisting of two stages, as shown in the Fig 3.9, namely: 1) pre-training a CNN regressor with hand-crafted features, and 2) fine tuning the pre-trained CNN regressor from stage 1 to classify breast masses. The first stage pre-trains the CNN model, which works as a regressor that approximates the values of hand-crafted features from the input image patch and segmentation mask. The idea behind this pre-training step with hand-crafted features lies with the fact that hand-crafted features have produced the state-of-the-art result in breast mass classification [4, 35, 44, 48, 50, 59, 82, 93, 94] and we want to integrate their importance in our classification model.

The hand-crafted features from a mammogram $\mathbf{x}$, given a bounding box $\mathbf{d}$ and segmentation mask $\mathbf{y}$, are extracted by applying the function

$$f_{\text{HF}}(\mathbf{x}, \mathbf{y}, \mathbf{d}) = \mathbf{z} \in \mathbb{R}^{H}, \tag{3.42}$$

where $f_{\text{HF}}(.)$ extracts a set of morphological, texture and intensity features as described in Sec. 3.4.3. We compute the texture and intensity features from the image patch localised by the bounding box $\mathbf{d}$ and morphological features from the segmentation mask $\mathbf{y}$.The pre-training of the CNN with hand-crafted features can be represented by the function

$$\mathbf{z}^{*} = f_{\text{CNNHF}}(\mathbf{x}, \mathbf{y}, \mathbf{d}, \theta_{\text{CNNHF}}), \tag{3.43}$$

where, f_{CNNHF} represents the CNN model with $L-1$ stages of convolutional, non-linear activation, max pooling and fully connected layer containing H nodes at the L^{th} stage, which outputs the approximated hand-crafted features, represented by $\mathbf{z}^{*}$. The training of this CNN regressor is done by minimising the following loss function:

$$\ell_{\text{HF}}(\theta_{\text{CNNHF}}) = \sum_{i=1}^{|\mathcal{D}|} \sum_{j}^{|\mathcal{A}_i|} ||\mathbf{z}^{*}_{(i,j)} - \mathbf{z}_{(i,j)}||_2, \tag{3.44}$$

where i is the index of the training images, j is the index for the mass in each training image i, $\mathbf{z}_{(i,j)}$ represents the vector of hand-crafted features from mass j and image i and $\mathbf{z}^{*}_{(i,j)}$ is the approximated hand-crafted features from the last layer of pre-trained CNN model. The second step of breast mass classification system, as shown in Fig. 3.9, adds another fully connected layer $L+1$ with softmax activation. We fine tune this CNN model by minimising the cross entropy loss using the class labels (benign or malignant) as:

$$\ell_{\text{FCNN}}(\theta_{\text{FCNN}}) = \sum_{i=1}^{|\mathcal{D}|} \sum_{j}^{|\mathcal{A}_i|} c_{(i,j)} \log \tilde{c}_{(i,j)} \tag{3.45}$$

where, θ_{FCNN} is the parameter of the fine tuned CNN model and $\tilde{c}$ is the class label predicted by this model.

3.7 Conclusions

In this chapter, we presented the general methodologies proposed in this thesis for the problem of breast mass detection, segmentation and classification. These methodologies are combined to form a fully automated CAD system for the analysis of breast masses in mammograms. These methodologies have been adapted for each problem being dealt in this thesis, which will be discussed in subsequent chapters.

Chapter 4

Deep Structured Learning for mass Segmentation from Mammograms

Neeraj Dhungel⋆**, Gustavo Carneiro**⋆**, Andrew P. Bradley**†

⋆ACVT, The University of Adelaide, Australia

†ITEE, The University of Queensland, Australia

The work contained in this chapter has been published in
IEEE International Conference on Image Processing (ICIP), 2015

The final publication is available at 10.1109/ICIP.2015.7351343)

Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. Deep structured learning for mass segmentation from mammograms. In *IEEE International Conference on Image Processing (ICIP)*, 2015.

Statement of Authorship

Title of Paper	Deep structured learning for mass segmentation from mammograms.
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. Deep structured learning for mass segmentation from mammograms. In *IEEE International Conference on Image Processing (ICIP)*, 2015.

Principal Author

Name of Principal Author (Candidate)	Neeraj Dhungel		
Contribution to the Paper	-Checked and processed database -Wrote all the coding -Built the conceptual idea -Wrote and refined the manuscript		
Overall percentage (%)	50%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	5th July 2016

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	-Built the conceptual idea -Wrote and refined the manuscript -Supervised the development of this work		
Signature		Date	5th July 2016

Name of Co-Author	Andrew P. Bradley		
Contribution to the Paper	-Refined the manuscript -Helped in the development of conceptual idea		
Signature		Date	5th July 2016

Please cut and paste additional co-author panels here as required.

Dhungel, N., Carneiro, G. & Bradley, A.P. (2015). Deep structured learning for mass segmentation from mammograms. In *IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, pp. 2950-2954.

Chapter 5

Tree Re-weighted belief propagation using deep learning potentials for mass segmentation from mammograms.

Neeraj Dhungel*, Gustavo Carneiro*, Andrew P. Bradley†

*ACVT, The University of Adelaide, Australia
†ITEE, The University of Queensland, Australia

The work contained in this chapter has been published in
IEEE International Symposium on Biomedical Imaging (ISBI), 2015.

The final publication is available at 10.1109/ISBI.2015.7163983

Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. Tree RE-weighted belief propagation using deep learning potentials for mass segmentation from mammograms. *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2015.

Statement of Authorship

Title of Paper	Tree RE-weighted belief propagation using deep learning potentials for mass segmentation from mammograms.
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. Tree RE-weighted belief propagation using deep learning potentials for mass segmentation from mammograms. *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2015.

Principal Author

Name of Principal Author (Candidate)	Neeraj Dhungel
Contribution to the Paper	-Checked and processed database -Wrote all the coding -Built the conceptual idea -Wrote and refined the manuscript
Overall percentage (%)	50%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date: 5th July 2016

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Gustavo Carneiro
Contribution to the Paper	-Built the conceptual idea -Wrote and refined the manuscript -Supervised the development of this work
Signature	Date: 5th July 2016

Name of Co-Author	Andrew P. Bradley
Contribution to the Paper	-Refined the manuscript -Helped in the development of conceptual idea
Signature	Date: 5/6/16

Please cut and paste additional co-author panels here as required.

Dhungel, N., Carneiro, G. & Bradley, A.P. (2015). Tree RE-weighted belief propagation using deep learning potentials for mass segmentation from mammograms. In *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, New York, NY, pp. 760-763.

Chapter 6

Deep learning and structured prediction for the segmentation of mass in mammograms

Neeraj Dhungel*, Gustavo Carneiro*, Andrew P. Bradley†

*ACVT, The University of Adelaide, Australia

†ITEE, The University of Queensland, Australia

The work contained in this chapter has been published in
18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

The final publication is available at 10.1007/978-3-319-24553-9_74

Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. Deep learning and structured prediction for the segmentation of mass in mammograms. *18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

Statement of Authorship

Title of Paper	Deep learning and structured prediction for the segmentation of mass in mammograms.
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. Deep learning and structured prediction for the segmentation of mass in mammograms. *18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

Principal Author

Name of Principal Author (Candidate)	Neeraj Dhungel		
Contribution to the Paper	-Checked and processed database -Wrote all the coding -Built the conceptual idea -Wrote and refined the manuscript		
Overall percentage (%)	50%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	5th July 2016

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	-Built the conceptual idea -Wrote and refined the manuscript -Supervised the development of this work		
Signature		Date	5th July 2016

Name of Co-Author	Andrew P. Bradley		
Contribution to the Paper	-Refined the manuscript -Helped in the development of conceptual idea		
Signature		Date	5/6/16

Please cut and paste additional co-author panels here as required.

Dhungel, N., Carneiro, G. & Bradley, A.P. (2015). Deep Learning and Structured Prediction for the Segmentation of Mass in Mammograms. In N. Navab, J. Hornegger, W. Wells & A. Frangi (Eds), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*. Lecture Notes in Computer Science, v. 9349. Springer, Cham.

Chapter 7

Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests

Neeraj Dhungel⋆, Gustavo Carneiro⋆, Andrew P. Bradley†

⋆ACVT, The University of Adelaide, Australia
†ITEE, The University of Queensland, Australia

The work contained in this chapter has been published in
International Conference on Digital Image Computing: Techniques and Applications (DICTA)
2015.

This paper was first presented in 3rd workshop on Breast Image Analysis (BIA) as part of 18th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI-BIA), Munich, Germany, which us gave the permission to publish this paper elsewhere.

The final publication is available at 10.1109/DICTA.2015.7371234

Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests. *International Conference on Digital Image Computing: Techniques and Applications (DICTA)* 2015.

Statement of Authorship

Title of Paper	Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests.
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests. *International Conference on Digital Image Computing: Techniques and Applications (DICTA)* 2015.

Principal Author

Name of Principal Author (Candidate)	Neeraj Dhungel		
Contribution to the Paper	-Checked and processed database -Wrote all the coding -Built the conceptual idea -Wrote and refined the manuscript		
Overall percentage (%)	50%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	5th July 2016

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	-Built the conceptual idea -Wrote and refined the manuscript -Supervised the development of this work		
Signature		Date	5th July 2016

Name of Co-Author	Andrew P. Bradley		
Contribution to the Paper	-Refined the manuscript -Helped in the development of conceptual idea		
Signature		Date	5/6/16

Please cut and paste additional co-author panels here as required.

Dhungel, N., Carneiro G. & Bradley, A.P. (2015). Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Adelaide, SA, pp. 1-8.

http://dx.doi.org/10.1109/DICTA.2015.7371234

Chapter 8

The Automated Learning of Deep Features for Breast Mass Classification from Mammograms

Neeraj Dhungel***, Gustavo Carneiro*****, Andrew P. Bradley**†

*ACVT, The University of Adelaide, Australia
†ITEE, The University of Queensland, Australia

The work contained in this chapter has been published in
Submitted to 19th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2016.

The final publication is under review

Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. The Automated Learning of Deep Features for Breast Mass Classification from Mammograms. *Submitted to 19th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016.

Statement of Authorship

Title of Paper	The Automated Learning of Deep Features for Breast Mass Classification from Mammograms.	
Publication Status	☐ Published	☑ Accepted for Publication
	☐ Submitted for Publication	☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. The Automated Learning of Deep Features for Breast Mass Classification from Mammograms. *19th International Confer- ence on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016.	

Principal Author

Name of Principal Author (Candidate)	Neeraj Dhungel		
Contribution to the Paper	-Checked and processed database -Wrote all the coding -Built the conceptual idea -Wrote and refined the manuscript		
Overall percentage (%)	50%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	5th July 2016

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	-Built the conceptual idea -Wrote and refined the manuscript -Supervised the development of this work		
Signature		Date	5th July 2016

Name of Co-Author	Andrew P. Bradley		
Contribution to the Paper	-Refined the manuscript -Helped in the development of conceptual idea		
Signature		Date	5/6/16

Please cut and paste additional co-author panels here as required.

The Automated Learning of Deep Features for Breast Mass Classification from Mammograms

Neeraj Dhungel[†] Gustavo Carneiro[†] Andrew P. Bradley[⋆] [⋆]

[†] ACVT, School of Computer Science, The University of Adelaide
[⋆] School of ITEE, The University of Queensland

Abstract. The classification of breast masses from mammograms into benign or malignant has been commonly addressed with machine learning classifiers that use as input a large set of hand-crafted features, usually based on general geometrical and texture information. In this paper, we propose a novel deep learning method that automatically learns features based directly on the optmisation of breast mass classification from mammograms, where we target an improved classification performance compared to the approach described above. The novelty of our approach lies in the two-step training process that involves a pre-training based on the learning of a regressor that estimates the values of a large set of hand-crafted features, followed by a fine-tuning stage that learns the breast mass classifier. Using the publicly available INbreast dataset, we show that the proposed method produces better classification results, compared with the machine learning model using hand-crafted features and with deep learning method trained directly for the classification stage without the pre-training stage. We also show that the proposed method produces the current state-of-the-art breast mass classification results for the INbreast dataset. Finally, we integrate the proposed classifier into a fully automated breast mass detection and segmentation, which shows promising results.

Keywords: deep learning, breast mass classification, mammograms

1 Introduction

Mammography represents the main imaging technique used for breast cancer screening [1] that uses the (mostly manual) analysis of lesions (i.e., masses and micro-calcifications) [2]. Although effective, this manual analysis has a trade-off between sensitivity (84%) and specificity (91%) that results in a relatively large number of unnecessary biopsies [3]. The main objective of computer aided diagnosis (CAD) systems in this problem is to act as a second reader with the goal of increasing the breast screening sensitivity and specificity [1]. Current automated mass classification approaches extract hand-crafted features from an image patch containing a breast mass, and subsequently use them in a classification process based on traditional machine learning methodologies, such as support vector machines (SVM) or multi-layer perceptron (MLP) [4]. One issue with this approach

[⋆] This work was partially supported by the Australian Research Council's Discovery Projects funding scheme (project DP140102794). Prof. Bradley is the recipient of an Australian Research Council Future Fellowship(FT110100623).

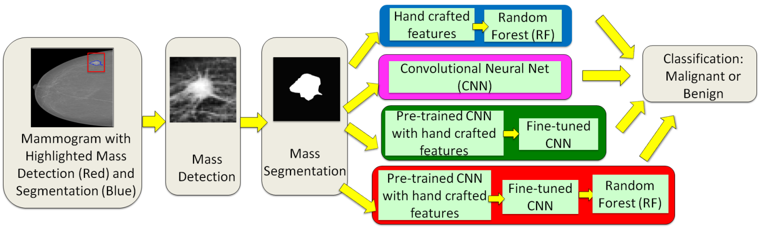

Mammogram with Highlighted Mass Detection (Red) and Segmentation (Blue)
Mass Detection
Mass Segmentation
Hand crafted features
Random Forest (RF)
Convolutional Neural Net (CNN)
Pre-trained CNN with hand crafted features
Fine-tuned CNN
Pre-trained CNN with hand crafted features
Fine-tuned CNN
Random Forest (RF)
Classification: Malignant or Benign

Fig. 1: Four classification models explored in this paper, where our main contribution consists of the last two models (highlighted in red and green).

is that the hand-crafted features are not optimised to work specifically for the breast mass classification problem. Another limitation of these methods is that the detection of image patches containing breast masses is typically a manual process [4, 5] that guarantees the presence of a mass for the segmentation and classification stages.

In this paper, we propose a new deep learning model [6, 7] which addresses the issue of producing features that are automatically learned for the breast mass classification problem. The main novelty of this model lies in the training stage that comprises two main steps: first stage acknowledges the importance of the aforementioned hand-crafted features by using them to pre-train our model, and the second stage fine-tunes the features learned in the first stage to become more specialised for the classification problem. We also propose a fully automated CAD system for analysing breast masses from mammograms, comprising a detection [8] and a segmentation [9] steps, followed by the proposed deep learning models that classify breast masses. We show that the features learned by our proposed models produce accurate classification results compared with the hand-crafted features [4, 5] and the features produced by a deep learning model without the pre-training stage [6, 7] (Fig. 1) using the INbreast [10] dataset. Also, our fully automated system is able to detect 90% of the masses at a 1 false positive per image, where the final classification accuracy reduces only by 5%.

2 Literature Review

Breast mass classification systems from mammograms comprise three steps: mass detection, segmentation and classification.The majority of classification methods still relies on the manual localisation of masses as their automated detection is still considered a challenging problem [4]. The segmentation is mostly an automated process generally based on active contour [11] or dynamic programming [4]. The classification usually relies on hand-crafted features, extracted from the detected image patches and their segmentation,which are fed into classifiers that classify masses into benign or malignant [4, 11, 5]. A common issue with these approaches is that they are tested on private datasets, preventing fair comparisons. A notable exception is the work by Domingues et al. [5] that uses the publicly available INbreast dataset [10]. Another issue is that the results from fully automated detection, segmentation and classification CAD systems are not (often) published in the open literature, which makes comparisons difficult.

Deep learning models have consistently shown to produce more accurate classification results compared to models based on hand-crafted features [6, 12]. Recently, these models have been successfully applied in mammogram classification [13], breast mass detection [8] and segmentation [9]. Carneiro et al. [13] have proposed a semi-automated mammogram classification using a deep learning model pre-trained with computer vision datasets, which differs from our proposal given that ours is fully automated and that we process each mass independently. Finally, for the fully automated CAD system, we use the deep learning models of detection [8] and segmentation [9] that produce the current state-of-the-art results on INbreast [10].

3 Methodology

Dataset The dataset is represented by $\mathcal{D} = \{(\mathbf{x}, \mathcal{A})_i\}_{i=1}^{|\mathcal{D}|}$, where mammograms are denoted by $\mathbf{x} : \Omega \rightarrow \mathbb{R}$ with $\Omega \in \mathbb{R}^2$, and the annotation for the $|\mathcal{A}_i|$ masses for mammogram i is represented by $\mathcal{A}_i = \{(\mathbf{d}, \mathbf{s}, c)_j\}_{j=1}^{|\mathcal{A}_i|}$, where $\mathbf{d}(i)_j = [x, y, w, h] \in \mathbb{R}^4$ represents the left-top position (x, y) and the width w and height h of the bounding box of the j^{th} mass of the i^{th} mammogram, $\mathbf{s}(i)_j : \Omega \rightarrow \{0, 1\}$ represents the segmentation map of the mass within the image patch defined by the bounding box $\mathbf{d}(i)_j$, and $c(i)_j \in \{0, 1\}$ denotes the class label of the mass that can be either benign (i.e., BI-RADS $\in \{1, 2, 3\}$) or malignant (i.e., BI-RADS $\in \{4, 5, 6\}$).

Classification Features The features are obtained by a function that takes a mammogram, the mass bounding box and segmentation, defined by:

$$f(\mathbf{x}, \mathbf{d}, \mathbf{s}) = \mathbf{z} \in \mathbb{R}^N. \tag{1}$$

In the case of **hand-crafted features**, the function $f(.)$ in (1) extracts a vector of morphological and texture features [4]. The morphological features are computed from the segmentation map $\mathbf{s}$ and consist of geometric information, such as area, perimeter, ratio of perimeter to area, circularity, rectangularity, etc. The texture features are computed from the image patch limited by the bounding box $\mathbf{d}$ and use the spatial gray level dependence (SGLD) matrix [4] in order to produce energy, correlation, entropy, inertia, inverse difference moment, sum average, sum variance, sum entropy, difference of average, difference of entropy, difference variance, etc. The hand-crafted features are denoted by $\mathbf{z}^{(H)} \in \mathbb{R}^N$.

The classification features from the **deep learning model** are obtained using a convolutional neural network (CNN) [7], which consists of multiple processing layers containing a convolution layer followed by a non-linear activation and a sub-sampling layer, where the last layers are represented by fully connected layers and a final regression/classification layer [7, 6]. Each convolution layer $l \in \{1, ..., L\}$ computes the output at location j from input at i using the filter $\mathbf{W}_m^{(l)}$ and bias $b_m^{(l)}$, where $m \in \{1, ..., M(l)\}$ denotes the number of features in layer l, as follows: $\widetilde{\mathbf{x}}^{(l+1)}(j) = \sigma(\sum_{i \in \Omega} \mathbf{x}^{(l)}(i) * \mathbf{W}_m^{(l)}(i, j) + b_m^{(l)}(j))$, where $\sigma(.)$ is

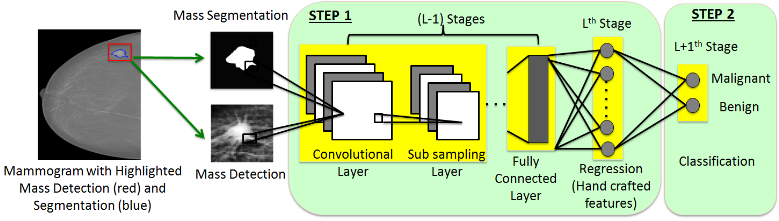


Fig. 2: Two steps of the proposed model with the pre-training of the CNN with the regression to the hand-crafted features (step 1), followed by the fine-tuning using the mass classification problem (step 2).

the activation function [7, 6], $\mathbf{x}^{(1)}$ is the original image, and $*$ is the convolution operator. The sub-sampling layer is computed by $\mathbf{x}^{(l)}(j) = \downarrow(\widetilde{\mathbf{x}}^{(l)}(j))$, where $\downarrow(.)$ is the subsampling function that pools the values (i.e., a max pooling operator) in the region $j \in \Omega$ of the input data $\widetilde{\mathbf{x}}^{(l)}(j)$. The fully connected layer is determined by the convolution equation above using a separate filter for each output location, using the whole input from the previous layer.

In general, the last layer of a CNN consists of a classification layer, represented by a softmax activation function. For our particular problem of mass classification, recall that we have a binary classification problem, defined by $c \in \{0, 1\}$ (Sec. 3), so the last layer contains two nodes (benign or malignant mass classification), with a softmax activation function [6]. The training of such a CNN is based on the minimisation of the regularised cross-entropy loss [6], where the regularisation is generally based on the ℓ_2 norm of the parameters θ of the CNN. In order to have a fair comparison between the hand-crafted and CNN features, the number of nodes in layer $L - 1$ must be N, which is the number of hand-crafted features in (1). It is well known that CNN can overfit the training data even with the regularisation of the weights and biases based on ℓ_2 norm, so a current topic of investigation is how to regularise the training more effectively [14].

One of the contributions of this paper is an experimental investigation of how to regularise the training for problems in medical image analysis that have traditionally used hand-crafted features. Our proposal is a two-step training process, where the first stage consists of training a regressor (see step1 in Fig. 2), where the output $\widetilde{\mathbf{x}}^{(L)}$ approximates the values of the hand-crafted features $\mathbf{z}^{(H)}$ using the following loss function:

$$J = \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{A}_i|} \|\mathbf{z}^{(H)}_{(i,j)} - \widetilde{\mathbf{x}}^{(L)}_{(i,j)}\|_2, \tag{2}$$

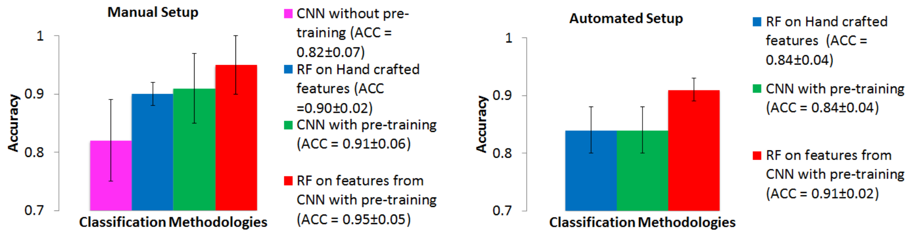
where i indexes the training images, j indexes the masses in each training image, and $\mathbf{z}^{(H)}_{(i,j)}$ denotes the vector of hand-crafted features from mass j and image i. This first step acts as a regulariser for the classifier that is sub-sequentially fine-tuned (see step 2 in Fig. 2).

Fully Automated Mass Detection, Segmentation and Classification The mass detection and segmentation methods are based on deep learning methods recently proposed by Dhungel et al. [8, 9]. More specifically, the detection consists of a cascade of increasingly more complex deep learning models, while the segmentation comprises a structured output model, containing deep learning potential functions. We use these particular methods given their use of deep learning methods (which facilitates the integration with the proposed classification), and their state-of-art performance on both problems.

4 Materials and Methods

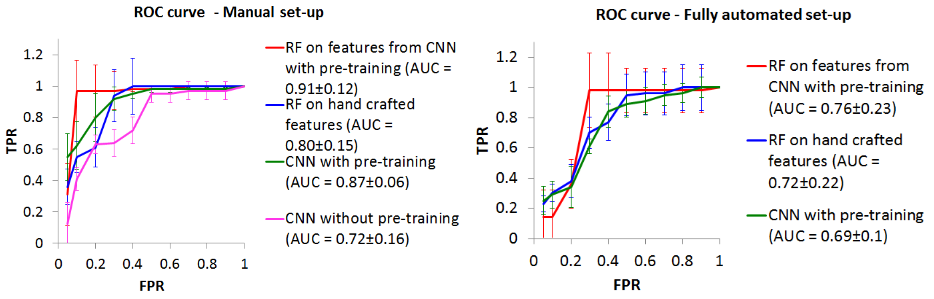
We use the publicly available INbreast dataset [10] that contains 115 cases with 410 images, where 116 images contain benign or malignant masses. Experiments are run using five fold cross validation by randomly dividing the 116 cases in a mutually exclusive manner, with 60% of the cases for training, 20% for validation and 20% for testing. We test our classification methods using a manual and an automated set-up, where the manual set-up uses the manual annotations for the mass bounding box and segmentation. The automated set-up first detects the mass bounding boxes [8] (we select a detection score threshold based on the training results that produces a TPR = 0.93 ± 0.05 and FPI = 0.8 on training data - this same threshold produces TPR of 0.90 ± 0.02 and FPI = 1.3 on testing data, where a detection is positive if the intersection over union ratio (IoU)>= 0.5 [8]). The resulting bounding boxes and segmentation maps are resized to 40 x 40 pixels using bicubic interpolation, where the image patches are contrast enhanced, as described in [11]. Then the bounding boxes are automatically segmented [9], where the segmentation results using only the TP detections has a Dice coefficient of 0.85 ± 0.01 in training and 0.85 ± 0.02 in testing. From these patches and segmentation maps, we extract 781 hand-crafted features [4] used to pre-train the CNN model and to train and test the baseline model using the random forest (RF) classifier [15].

The CNN model for step 1 (pre-training in Fig. 2) has an input with two channels containing the image patch with a mass and respective segmentation mask; layer 1 has 20 filters of size 5×5, followed by a max-pooling layer (subsamples by 2); layer 2 contains 50 filters of size 5×5 and a max-pooling that subsamples by 2; layer 3 has 100 filters of size 4×4 followed by a rectified linear unit (ReLU) [16]; layer 4 has 781 filters of size 4x4 followed by a ReLU unit; layer 5 comprises a fully-connected layer of 781 nodes that is trained to approximate the hand-crafted features, as in (2). The CNN model for step 2 (fine-tuning in Fig. 2) uses the pre-trained model from step 1, where a softmax layer containing two nodes (representing the benign versus malignant classification) is added, and the fully-connected layers are trained with drop-out of 0.3 [14]. Note that for comparison purposes, we also train a CNN model without the pre-training step to show its influence in the classification accuracy. In order to improve the regularisation of the CNN models, we artificially augment by 10-fold the training data using geometric transformations (rotation, translation and scale). Moreover, using the hand-crafted features, we train an RF classifier [15], where model selection is performed using the validation set of each cross validation training set. We also train a RF classifier using the 781 features from the second last fully-connected layer of the fine-tuned CNN model. We carried out all our experiments



(a) Manual set-up (b) Automated set-up

Fig. 3: Accuracy on test data of the methodologies explored in this paper.



(a) Manual set-up (b) Automated set-up

Fig. 4: ROC curves of various methodologies explored in this paper on test data.

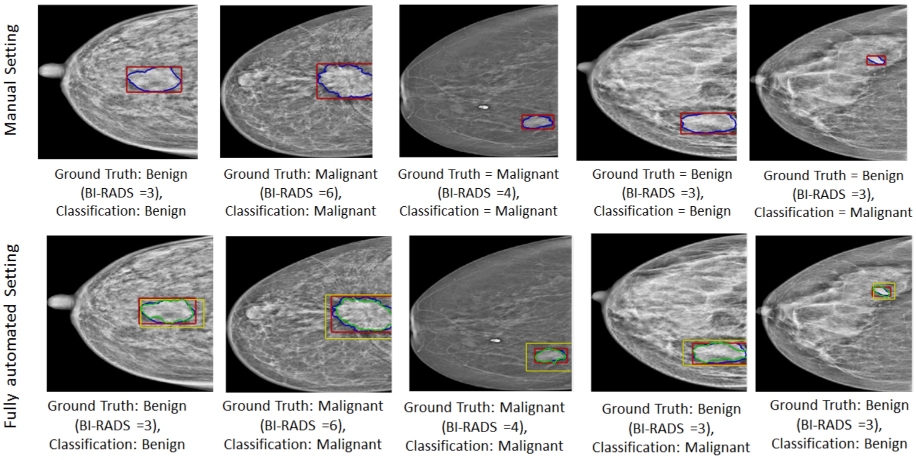
using a computer with the following configuration: Intel(R) Core(TM) i5-2500k 3.30GHz CPU with 8GB RAM and graphics card NVIDIA GeForce GTX 460 SE 4045 MB. We compare the results of the methods explored in this paper with receiver operating characteristic (ROC) curve and classification accuracy (ACC).

5 Results

Figures 3(a-b) show a comparison amongst the models explored in this paper using classification accuracy for both manual and automated set-ups. The most accurate model in both set-ups is the RF on features from the CNN with pre-training with ACC of 0.95 ± 0.05 on manual and 0.91 ± 0.02 on automated set-up (results obtained on test set). Similarly, Fig. 4(a-b) display the ROC curves that also show that RF on features from the CNN with pre-training produces the best overall result with the area under curve (AUC) value of 0.91 ± 0.12 for manual and 0.76 ± 0.23 for automated set-up on test sets. In Tab. 1, we compare our results with the current state-of-the-art techniques in terms of accuracy (ACC), where the second column describes the dataset used and whether it can be reproduced (‘Rep’) because it uses a publicly available dataset, and the third

Table 1: Comparison of the proposed and state-of-the-art methods on test sets.

Methodology	Dataset (Rep?)	set-up	ACC
Proposed RF on CNN with pre-training	INbreast (Yes)	Manual	0.95 ± 0.05
Proposed CNN with pre-training	INbreast (Yes)	Manual	0.91 ± 0.06
Proposed RF on CNN with pre-training	INbreast(Yes)	Fully automated	0.91 ± 0.02
Proposed CNN with pre-training	INbreast (Yes)	Fully automated	0.84 ± 0.04
Domingues et. al [5]	INbreast (Yes)	Manual	0.89
Varela et. al [4]	DDSM (No)	Semi-automated	0.81
Ball et. al [11]	DDSM (No)	Semi-automated	0.87



Ground Truth: Benign (BI-RADS =3), Classification: Benign

Ground Truth: Malignant (BI-RADS =6), Classification: Malignant

Ground Truth = Malignant (BI-RADS =4), Classification = Malignant

Ground Truth = Benign (BI-RADS =3), Classification = Benign

Ground Truth = Benign (BI-RADS =3), Classification = Malignant

Ground Truth: Benign (BI-RADS =3), Classification: Benign

Ground Truth: Malignant (BI-RADS =6), Classification: Malignant

Ground Truth: Malignant (BI-RADS =4), Classification: Malignant

Ground Truth: Benign (BI-RADS =3), Classification: Malignant

Ground Truth: Benign (BI-RADS =3), Classification: Benign

Fig. 5: Results of RF on features from the CNN with pre-training on test set. Red and blue lines denote manual detection and segmentation whereas yellow and green lines are the automated detection and segmentation.

column, denoted by ‘set-up’, describes the method of mass detection and segmentation (semi-automated means that detection is manual, but segmentation is automated). The running time for the fully automated system is 41 s, divided into 39 s for the detection, 0.2 s for the segmentation and 0.8 s for classification. The training time for classification is 6 h for pre-training, 3 h for fine-tuning and 30 m for the RF classifier training.

6 Discussion and Conclusions

The results from Figures 3 and 4 (both manual and automated set-ups) show that the CNN model with pre-training and RF on features from the CNN with pre-training are better than the RF on hand-crafted features and CNN without pre-training. Another important observation from Fig. 3 is that the RF classifier performs better than CNN classifier on features from CNN with pre-training. The results for the CNN model without pre-training in automated set-up are not shown because they are not competitive, which is expected given its relatively worse performance in the manual set-up. In order to verify the statistical significance of these results, we perform the Wilcoxon paired signed-rank test between the RF on hand-crafted features and RF on features from the CNN with pre-training, where the p-value obtained is 0.02, which indicates that the result is significant (assuming 5% significance level). In addition, both the proposed CNN with pre-training and RF on features from CNN with pre-training generalise well, where the training accuracy in the manual set-up for the former is 0.93 ± 0.06 and the latter is 0.94 ± 0.03.

In this paper we show that the proposed two-step training process involving a pre-training based on the learning of a regressor that estimates the values of a large set of hand-crafted features, followed by a fine-tuning stage that learns the breast mass classifier produces the current state-of-the-art breast mass classification results on INbreast. Finally, we also show promising results from a fully automated breast mass detection, segmentation and classification system.

References

1. Giger, M.L., Karssemeijer, N., Schnabel, J.A.: Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. Annual review of biomedical engineering **15** (2013) 327–357
2. Fenton, J.J., Taplin, S.H., Carney, P.A., et al.: Influence of computer-aided detection on performance of screening mammography. New England Journal of Medicine **356**(14) (2007) 1399–1409
3. Elmore, J.G., Jackson, S.L., Abraham, L., et al.: Variability in interpretive performance at screening mammography and radiologists characteristics associated with accuracy1. Radiology **253**(3) (2009) 641–651
4. Varela, C., Timp, S., Karssemeijer, N.: Use of border information in the classification of mammographic masses. Physics in Medicine and Biology **51**(2) (2006)
5. Domingues, I., Sales, E., Cardoso, J., Pereira, W.: Inbreast-database masses characterization. XXIII CBEB (2012)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. Volume 1. (2012)
7. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks **3361** (1995)
8. Dhungel, N., Carneiro, G., Bradley, A.: Automated mass detection in mammograms using cascaded deep learning and random forests. In: DICTA. (Nov 2015)
9. Dhungel, N., Carneiro, G., Bradley, A.P.: Deep learning and structured prediction for the segmentation of mass in mammograms. In: MICCAI. Springer (2015)
10. Moreira, I.C., Amaral, I., Domingues, I., et al.: Inbreast: toward a full-field digital mammographic database. Academic Radiology **19**(2) (2012) 236–248
11. Ball, J.E., Bruce, L.M.: Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation. In: EMBS 2007, IEEE (2007)
12. Farabet, C., Couprie, C., Najman, L., et al.: Learning hierarchical features for scene labeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35**(8) (2013)
13. Carneiro, G., Nascimento, J., Bradley, A.P.: Unregistered multiview mammogram analysis with pre-trained deep learning models. In: MICCAI. Springer (2015)
14. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1) (2014) 1929–1958
15. Breiman, L.: Random forests. Machine learning **45**(1) (2001) 5–32
16. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML-10. (2010)

Chapter 9

A Deep Learning approach to fully automated analysis of Masses in Mammograms

Neeraj Dhungel⋆, Gustavo Carneiro⋆, Andrew P. Bradley†

⋆ACVT, The University of Adelaide, Australia
†ITEE, The University of Queensland, Australia

The work contained in this chapter has been published in
Submitted to Medical Image Analysis (MedIA), 2016.

The final publication is under review

Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. A Deep Learning approach to fully automated analysis of Masses in Mammograms. *Submitted to Medical Image Analysis (MedIA)*

Statement of Authorship

Title of Paper	A Deep Learning approach to fully automated analysis of Masses in Mammograms.
Publication Status	□ Published □ Accepted for Publication ☑ Submitted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Neeraj Dhungel, Gustavo Carneiro, Andrew P. Bradley. A Deep Learning approach to fully automated analysis of Masses in Mammograms. *Submitted to Medical Image Anal- ysis (MedIA)* .

Principal Author

Name of Principal Author (Candidate)	Neeraj Dhungel		
Contribution to the Paper	-Checked and processed database -Wrote all the coding -Built the conceptual idea -Wrote and refined the manuscript		
Overall percentage (%)	50%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	5th July 2016

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	-Built the conceptual idea -Wrote and refined the manuscript -Supervised the development of this work		
Signature		Date	5th July 2016

Name of Co-Author	Andrew P. Bradley		
Contribution to the Paper	-Refined the manuscript -Helped in the development of conceptual idea		
Signature		Date	5/6/16

A Deep Learning Approach to Fully Automated Analysis of Masses in Mammograms

Neeraj Dhungel[a,*], Gustavo Carneiro[b], Andrew P. Bradley[c,1]

[a] *Australian Centre for Visual Technologies, The University of Adelaide, Australia*
[b] *Australian Centre for Visual Technologies, The University of Adelaide, Australia*
[c] *ITEE, The University of Queensland, Australia*

Abstract

We present a fully automated method for the problem of detecting, segmenting and classifying breast masses from mammograms. This is a long standing problem due to low signal-to-noise ratio in the visualisation of breast masses, combined with their large variability in terms of shape, size, appearance and location. We break the problem down into three stages: mass detection, mass segmentation, and mass classification. For the detection, we propose a cascade of deep learning methods to select hypotheses that are refined based on Bayesian optimisation. For the segmentation, we propose the use of deep structured output learning that is subsequently refined by a level set method. Finally, for the classification, we propose the use of a deep learning classifier, which is pre-trained with a regression to hand-crafted feature values and fine-tuned based on the annotations of the breast mass classification dataset. We test our proposed system on the publicly available INbreast dataset and compare the results with the current state-of-the-art methodologies. This evaluation shows that our fully automated system detects 90% of masses at 1 false positive per image, has a segmentation accuracy of around 0.85 (Dice index), and overall classifies masses as malignant or benign with sensitivity (Se) of 0.98 and specificity (Sp) of 0.7.

[*] Corresponding author
Email address: `neeraj.dhungel@adelaide.edu.au` (Neeraj Dhungel)

[1] This work was partially supported by the Australian Research Council's Discovery Projects funding scheme (project DP140102794). Prof. Bradley is the recipient of an Australian Research Council Future Fellowship (FT110100623).

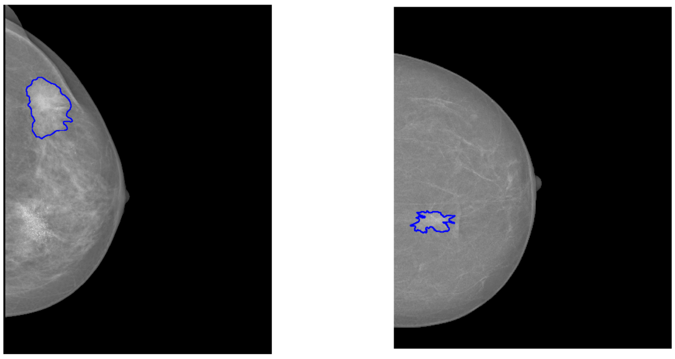
Preprint submitted to Journal of Medical Image Analysis *June 29, 2016*

Keywords: Mammograms, Masses, Detection, Segmentation, Classification, Deep Learning, Bayesian Optimisation, Transfer Learning, Structured Output Learning

1. Introduction

Breast cancer is one of the major diseases affecting the lives of many women worldwide. Statistical data published by the World Health Organisation (WHO) show that 23% of all cancer related cases and 14% of cancer related deaths amongst women are due to breast cancer (Jemal et al. (2008)). One of the most effective ways to reduce breast cancer mortality and morbidity is with breast screening programs that use mammograms as the main imaging modality (AIHW (2012)) (see Fig. 1). In these programs, the analysis of breast masses from mammograms represents an important task in the diagnosis of breast cancer, which is mostly a manual process that is susceptible to the subjective assessment of a clinical expert. Recent studies by Dromain et al. (2013) and Elmore et al. (2009) show that this manual analysis has a sensitivity of 84% and a specificity of 91% in the diagnosis of breast cancer (Giger and Pritzker (2014)). The classification accuracy of this manual interpretation can be improved with the use of a second reading of the mammogram by another clinical expert or by a computer-aided diagnosis (CAD) system (Giger and Pritzker (2014)). However, such CAD systems must be robust to false positives and false negatives to be useful in a clinical setting.

One of the main tasks performed by a CAD system is the detection, segmentation and classification of breast masses. This is challenging task due to the low signal-to-noise ratio of the mass visualisation, combined with the lack of consistent patterns of shape, size, appearance and location of breast masses (Oliver et al. (2010); Tang et al. (2009)). Furthermore, the relatively low availability of annotated datasets containing full field digital mammograms (FFDM), the most common type of breast imaging used in the field (see Fig. 1), hinders the development and evaluation of CAD systems. Current methodologies for mass



(a) Malignant mass (b) Benign mass

Figure 1: Two types of breast mass of a full field digital mammogram (FFDM) from the INbreast dataset (Moreira et al. (2012)): a) benign and b) malignant.

detection usually rely on a candidate region detection that uses several filters such as morphological, difference of Gaussian, Laplacian of Gaussian, etc. (Kozegar et al. (2013); Beller et al. (2005); te Brake et al. (2000); Campanini et al. (2004); Eltonsy et al. (2007); Sampat et al. (2008); Bellotti et al. (2006); Wei et al. (2005)). The detected candidates are then pruned using the responses of different types of classifiers, such as support vector machines (SVM), linear discriminant analysis (LDA) or artificial neural network (ANN) (Kozegar et al. (2013); Beller et al. (2005); te Brake et al. (2000); Campanini et al. (2004); Eltonsy et al. (2007); Sampat et al. (2008); Bellotti et al. (2006); Wei et al. (2005)). The main drawbacks of such mass detection methods are that they can generate a large number of false positives, while missing a good proportion of true positives (Oliver et al. (2010)), and the detected bounding boxes are often not accurately aligned with the mass, which can have a negative impact on the subsequent segmentation and classification stages. Segmentation methods generally work by taking the bounding boxes from the detection stage and segmenting them based on shape and appearance models using graph-based or level set methods (Rahmati et al. (2012); Cardoso et al. (2015)). Here the main challenges are related to the robustness of these shape and appearance models, the optimality of the proposed solution and the run-time/memory complexity of the method. Finally, mass classification typically uses hand-crafted features, extracted from the bounding boxes and segmentation maps, with traditional classification approaches, based on SVM or ANN (Varela et al. (2006); Shi et al. (2008); Domingues et al. (2012)). The main limitation of these mass classification approaches lies in the lack of optimality and complex design and selection of discriminatory hand-crafted features.

This paper is an extension of our previous works on mass detection (Dhungel et al. (2015a)), segmentation (Dhungel et al. (2015b)), and classification (Dhungel et al. (2016)) (see Fig. 2). Our previous work on mass detection (Dhungel et al. (2015a)) is based on multi-scale deep belief nets (m-DBN) and Gaussian mixture model (GMM), which is followed by a false positive reduction step based on the classification results provided by a convolutional neural network (CNN) and a random forest classifier (RF). In this paper, we extend our previous mass detection approach (Dhungel et al. (2015a)) with a more precise alignment of the bounding box with respect to the breast mass based on Bayesian optimisation (Zhang et al. (2015)). Moreover, our proposed mass segmentation methodology (Dhungel et al. (2015b)) is represented by a graph-based model that relies on unary potential functions based on deep learning methods (Dhungel et al. (2015b,c,d)). Parameter learning in the proposed graph-based approach is based on truncated fitting (Domke (2013)), while inference is performed with tree reweighted belief propagation (TRW) (Wainwright et al. (2003); Domke (2013)). The main novelties introduced in this paper, compared to our previous works on segmentation (Dhungel et al. (2015b,a)), is the use of the automated mass detection (Dhungel et al. (2015a)), replacing the manual mass detection, and a refinement stage based on level set methods (Chan et al. (2001)). Finally, the classification stage, based on deep learning methods, takes the appearance and shape from the automatically detected and segmented bounding boxes and produces the final mass classification (Dhungel et al. (2016)). The interesting aspect of this classification stage lies in our transfer learning approach: we pre-train a deep learning regressor to approximate the values produced by hand-crafted

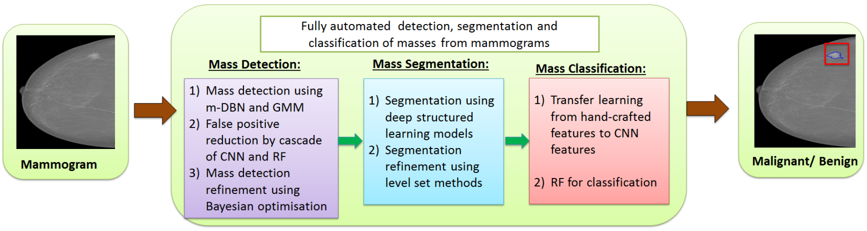

Mammogram
Fully automated detection, segmentation and classification of masses from mammograms
Mass Detection:
1) Mass detection using m-DBN and GMM
2) False positive reduction by cascade of CNN and RF
3) Mass detection refinement using Bayesian optimisation
Mass Segmentation:
1) Segmentation using deep structured learning models
2) Segmentation refinement using level set methods
Mass Classification:
1) Transfer learning from hand-crafted features to CNN features
2) RF for classification
Malignant/ Benign

Figure 2: Our proposed methodology of breast mass detection, segmentation and classification. Mass detection is done using mass candidate generation and false positive reduction (Dhungel et al. (2015a)) with a new detection refinement. Segmentation is carried out using our previously proposed work on deep structured learning (Dhungel et al. (2015b)), which is followed by a segmentation refinement step. Finally, classification is reached by training a CNN in two steps, where the first step is a regressor that estimates hand-crafted features followed by a second step that fine-tunes the model based on the mass classification problem.

features (Varela et al. (2006)), the network is then fine-tuned based on the mass classification problem to improve overall classification accuracy.

The detection, segmentation and classification accuracy produced by our fully automated breast analysis are measured on the publicly available INbreast dataset (Moreira et al. (2012)), which is the largest publicly available dataset of annotated FFDM mammograms in the field. This dataset contains 410 FFDM mammograms of the left and right breasts from 115 patients from two views: cranio-caudal (CC) and medio-lateral oblique (MLO). The accuracy of the automated mass detection, segmentation and classification system is compared to the manual annotations using the following measures: the free response operating characteristic (FROC) curve, average precision curve, pixel based true positive rate, Dice index, classification accuracy, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC). The results show that our system for automated detection, segmentation and classification of breast masses correlates well with the ground truth annotations. The results also show that our approach has results for each stage that are better than the current state-of-the-art methods. The final results from our fully automated system show that it is able to detect 90% of masses at one false positive rate per image, with segmentation accuracy of 85%, where the final classification (into benign or malignant) for the detected masses reaches sensitivity (Se) of 0.98 and specificity (Sp) of 0.7.

2. Literature Review

In this section, we review the literature for the problems of mass detection, segmentation and classification in mammograms. We also discuss the current deep learning methods that are relevant to our work.

Systems that can analyse mammograms depend heavily on the detection of breast masses, which is a challenging problem that, to a large extent, has not been fully solved (Fenton et al. (2007)). Several methodologies have been proposed for this problem, usually consisting of two stages: candidate mass detection by relatively simple image filters, followed by a false positive pruning stage (Kozegar et al. (2013); Beller et al. (2005); te Brake et al. (2000); Campanini et al. (2004); Eltonsy et al. (2007); Sampat et al. (2008); Bellotti et al. (2006); Wei et al. (2005)). The detection accuracy of these methods tends to be relatively poor due to the low capacity of the proposed models that does not allow a robust modelling of the shape, size and intensity variations of masses. In addition, most of the previously proposed methods have been tested on datasets that are not publicly available, which makes the comparison between methods an impossible task. Therefore, we propose the use of high capacity deep learning models (Girshick et al. (2014)) with the INbreast dataset (Moreira et al. (2012)) that is publicly available and contains high quality FFDM mammograms and precise expert annotations. We also propose the use of a detection refinement step (Zhang et al. (2015)) that improves the precision of the mass detection - a step that is not generally found in previous works.

The mass segmentation step is generally present in breast mass analysis systems because of the association between mass shape irregularities and the probability of cancer (Giger and Pritzker (2014)). It is important to note that mass segmentation is a step that is not explicitly undertaken in regular manual breast screening exams, and for that reason, it is difficult to acquire expert annotations. This means that annotated datasets tend to have a limited number of a training samples for that particular problem, which makes the design of a robust mass segmentation algorithm a challenging task. In spite of that, there have been a large number of methods proposed, such as the ones based on Markov random field models, with optimal inference but sub-optimal training (Cardoso et al. (2015); Rojas Domínguez and Nandi (2009); Song et al. (2009); Timp and Karssemeijer (2004); Yu et al. (2012)), level set methods with sub-optimal training and inference with strong shape priors (Ball and Bruce (2007); Rahmati et al. (2012); Sahiner et al. (2001); Sethian (1999); Shi et al. (2007); te Brake et al. (2000)). The main issues with the majority of mass segmentation methods are that they are evaluated on manually detected masses, are based on sub-optimal training or inference algorithms, and use training/testing datasets that are not publicly available. Our proposed mass segmentation methodology (Dhungel et al. (2015b)) uses structured prediction models based on hierarchical deep learning potential functions, producing optimal training and inference procedures (Dhungel et al. (2015b)). It also uses the results from our proposed automated mass detection method introduced above and relies on the publicly available INbreast dataset (Moreira et al. (2012)). Furthermore, we propose a segmentation refinement stage, based on a level set method (Chan et al. (2001)), that adjusts the delineation to the high-resolution input image - this stage is also not generally found in previous papers.

Breast mass classification is usually a semi-automated process that uses a set of hand-crafted features based on morphological features describing the geometrical structure of mass, and texture features computed from the intensity distribution of mass (Varela et al. (2006); Ball and Bruce (2007); Domingues et al. (2012)). These features are then used as the input to traditional machine learning classifiers, such as support vector machine (SVM) and artificial neural network (ANN), to classify masses into malignant or benign (Varela et al. (2006); Ball and Bruce (2007); Domingues et al. (2012)). Similarly to the mass segmentation problem presented above, mass classification methods (Varela et al. (2006); Ball and Bruce (2007)) usually use datasets that are not publicly available and depend on manually detected and segmented masses. In contrast, our proposed mass classification relies on automatically detected and segmented masses and uses the publicly available INbreast dataset (Moreira et al. (2012)). Furthermore, we explore deep learning models for this task which in principle can learn features directly from the input mass image and segmentation, but the robustness of this learning process is related to the size of the annotated training set. Given that the INbreast dataset does not contain a large annotated training set, we explore a pre-training process that regresses the results of hand-crafted features (Varela et al. (2006)), which is followed by a fine-tuning process that trains a classifier using the INbreast dataset annotations.

In computer vision, deep learning models have consistently been shown to produce more accurate classification results (e.g., object detection, semantic segmentation and classification) compared to previously proposed machine learning models (LeCun and Bengio (1995); Krizhevsky et al. (2012); Farabet et al. (2013); Girshick et al. (2014); Zhang et al. (2015)). A particularly interesting advantage of deep learning models is their ability to automatically learn a rich hierarchy of features for complex classification problems, avoiding problems associated with the hand-crafting of features: feature set sub-optimality, and complexity of the feature designing and selection process. This motivated us to explore deep learning as underlying framework for fully automated analysis (detection, segmentation and classification) of masses from mammograms. Also, the detected and segmented masses can be displayed to aid expert interpretation of our CAD system's decisions. Nevertheless, the deep learning models proposed in computer vision, containing several large annotated datasets, must be adapted to the medical imaging domain that has much smaller annotated datasets. This adaptation includes the use of pre-trained models (Carneiro et al. (2015)), an increase in the number of training images (Cireşan et al. (2013)), or a combination with other machine learning techniques (Dhungel et al. (2015a,b); Ngo and Carneiro (2014)). In this paper, we explore the first and the last ideas above, i.e., pre-trained models and the combination with other machine learning methods (Dhungel et al. (2016)).

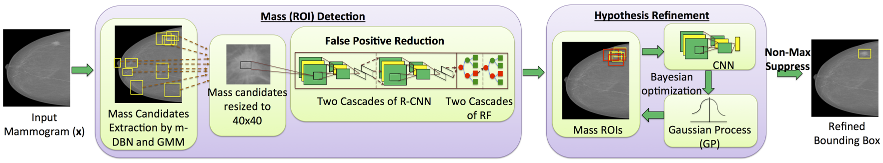


Figure 3: The proposed mass detection consists of two stages of mass ROI detection followed by hypothesis refinement. The Mass ROI detection is based on the results of m-DBN and GMM to generate candidates, followed by a false positive reduction using cascades of CNN and RF; and the hypothesis refinement is based on Bayesian optimisation.

3. Methodology

In this section, we first define the dataset used to train and test the proposed system, then we explain each stage of mass detection, segmentation and classification.

3.1. Dataset

The annotated dataset is represented by $\mathcal{D} = \{(\mathbf{x}, \mathcal{A})_i\}_{i=1}^{|\mathcal{D}|}$, where mammograms are denoted by $\mathbf{x} : \Omega \rightarrow \mathbb{R}$ with $\Omega \in \mathbb{R}^2$, and the annotation for the $|\mathcal{A}_i|$ masses for mammogram i is represented by $\mathcal{A}_i = \{(\mathbf{d}, \mathbf{y}, c)_j\}_{j=1}^{|\mathcal{A}_i|}$, where $\mathbf{d}_{i,j} = [x, y, w, h] \in \mathbb{R}^4$ represents the left-top position (x, y) and the width w and height h of the bounding box of the j^{th} mass of the i^{th} mammogram, $\mathbf{y}_{i,j} : \Omega \rightarrow \{0, 1\}$ represents the segmentation map of the mass within the image patch defined by the bounding box $\mathbf{d}_{i,j}$ and $c_{i,j} \in \{0, 1\}$ denotes the class label of the mass that can be either benign(i.e., BI-RADS $\in \{1, 2, 3\}$) or malignant (i.e., BI-RADS $\in \{4, 5, 6\}$).

3.2. Mass Detection

As depicted in Figure 3, our mass detection algorithm (Dhungel et al. (2015a)) consists of a cascade of classifiers, where the main goal of each stage is to keep the true positive detections while reducing the proportion of false positive detections and then improve the precision of bounding box detection. This requires classifiers with relative small memory and run-time complexities in the first stages to eliminate the "obvious" false positives. Then the later stage classifiers increase in complexity in order to be able to handle the more difficult candidates containing the true positives and not so obvious false positives. After finding the mass candidates, their localisation and scale still need to be refined in order to help the next stages of the system: the mass segmentation and classification.

3.2.1. Mass ROI Detection

The first stage of the detection consists of the generation of a set of N_{RGH} mass candidates, comprising their bounding boxes $\{\mathbf{d}_n^*\}_{n=1}^{N_{\text{RGH}}}$ and rough segmentation masks $\{\widetilde{\mathbf{y}}_n^*\}_{n=1}^{N_{\text{RGH}}}$ for a mammogram $\mathbf{x}$, defined by

$$\{\mathbf{d}_n^*, \widetilde{\mathbf{y}}_n^*\}_{n=1}^{N_{\text{RGH}}} = f_{\text{RGH}}(\mathbf{x}, \theta_{\text{ROI}}), \tag{1}$$

where $f_{\text{RGH}}(.)$ is a model defined by parameters θ_{RGH}. This function works by combining the detection results of a coarse-to-fine deep belief network (m-DBN) model and of a Gaussian mixture model (GMM). The m-DBN model uses a grid search on a coarse resolution of image $\mathbf{x}$, where each grid point is classified into positive or negative based on a square input of fixed size $S \times S$ extracted from around that grid point, and the output is represented by a softmax activation function. Then all points classified as positives are passed on to the next finer resolution stage to be classified in a similar manner - this process repeats for three coarse to fine stages, where the image resolution increases steadily between each stage. The training of this DBN (Hinton et al. (2006)) at each resolution level uses a training set of positive patches extracted from the grid points (a positive patch is defined by the central point that belongs to an annotated mass) and negative patches from the detection of previous stage, where the first stage uses randomly sampled negative patches (a negative patch is defined by a central point that does not belong to an annotated mass). The GMM (Dhungel et al. (2015a)) model works only on the finest image resolution with a pixel-wise classification, and this model is trained from the annotated training samples in order to estimate the likelihood that a pixel grey value represents part of a breast mass, or background. Note that this GMM model will produce a posterior probability that needs to be thresholded to produce the final estimated positive and negative labels, where this threshold varies from 0.3 to 0.9. The pixel-wise classification from m-DBN and GMM are then joined with a union operator, where a connected component analysis identifies the N_{RGH} mass candidates in (1).

False positives amongst the generated mass candidates in $\{\mathbf{d}_n^*, \widetilde{\mathbf{y}}_n^*\}_{n=1}^{N_{\text{RGH}}}$ are then pruned by a cascade of R-CNNs (Girshick et al. (2014); Dhungel et al. (2015a)), which extracts the features from the last layer of a CNN model and classifies it using a linear SVM (Cortes and Vapnik (1995)). A CNN (LeCun and Bengio (1995); Krizhevsky et al. (2012)) model consists of multiple processing stages, with each stage comprising two layers: linear filtering from the convolutional layer that generates responses, which are transformed via a non-linear activation function, and the pooling and sub-sampling layer that reduces the data size for the next stage. The CNN model has a final stage that consists of a fully connected layer (LeCun and Bengio (1995); Krizhevsky et al. (2012)). Each R-CNN stage is represented by:

$$\{\mathbf{d}_n^*, \widetilde{\mathbf{y}}_n^*\}_{n=1}^{N_{\text{RCNN}}} = f_{\text{RCNN}}(\mathbf{x}, \{\mathbf{d}_n^*, \widetilde{\mathbf{y}}_n^*\}_{n=1}^{N_{\text{RGH}}}, \theta_{\text{RCNN}}), \tag{2}$$

where $f_{\text{RCNN}}(.)$ is a model defined by parameters θ_{RCNN} (the weights and biases of the CNN and the linear SVM parameters), and $N_{\text{RCNN}} \leq N_{\text{RGH}}$ (i.e., the number of candidates tends to reduce after the R-CNN stage). The input for the R-CNN model in (2) is defined by taking each bounding box $\mathbf{d}_n^*$ and extracting an image patch from $\mathbf{x}$, which is then resized to $M \times M$ using bi-cubic interpolation and contrast enhanced (Ball and Bruce (2007)). The training of the CNN involves taking the N_{RGH} candidates and define a set of positive and negative samples, by looking at the overlap between the estimated and annotated bounding boxes, and the objective of this training is to minimise a softmax classification loss. Specifically, if the overlap is bigger than 0.2, then it represents a positive sample, otherwise, it is a negative sample. Instead of using this classification result from the CNN, we notice that by taking a feature vector built from the last fully-connected layer (before the the softmax layer), and use it in a linear SVM classifier, we are able to produce more accurate classification results. All candidates that survived the first cascade of the R-CNN are then passed through to the second cascade of R-CNN to further reduce the number of false positive detections (Dhungel et al. (2015a)).

After the R-CNN stage, we still have a relatively high false positive rate and as a result a new round of classifiers needs to be introduced. Note that at this stage, the classification problem is complex, so we need a high capacity model that can learn to represent this classification problem. Therefore, we first extract a large number of hand-crafted features extracted from the masses candidate of the second stage $\{\mathbf{d}^*_n, \widetilde{\mathbf{y}}^*_n\}_{n=1}^{J_{\text{RCNN}}}$ and feed them to a cascade of random forest (RF) classifiers (Breiman (2001)). In particular, we use object based morphological features such as number of perimeter pixels, area, perimeter-to-area ratio, circularity, rectangularity, and five normalised radial length (NRL) features (Wei et al. (2005); Dhungel et al. (2015a)), in addition to the texture features from grey level co-occurrence matrix (GLCM) (Wei et al. (2005); Dhungel et al. (2015a)). In total, we have 781 hand-crafted features available at this stage. The RF classifier is defined by

$$\{\mathbf{d}^*_n, \widetilde{\mathbf{y}}^*_n\}_{n=1}^{N} = f_{\text{RF}}(\mathbf{x}, \{\mathbf{d}^*_n, \widetilde{\mathbf{y}}^*_n\}_{n=1}^{N_{\text{RCNN}}}, \theta_{\text{RF}}), \tag{3}$$

where $f_{\text{RF}}(.)$ represents a random forest classifier defined by parameters θ_{RF} (number of trees, number of leaves in each tree, etc.), and $N \leq N_{\text{RCNN}}$ (i.e., the number of candidates tends to be smaller after the RF stage).

3.3. Hypothesis Refinement

This hypothesis refinement step is one of the novel contributions of this paper, where the objective is the adjustment of the bounding boxes in the set $\{\mathbf{d}^*_n, \widetilde{\mathbf{y}}^*_n\}_{n=1}^{N}$, produced by the RF classifier in (3), such that they fit more tightly around the detected breast masses. Assuming that we have a scoring function defined by

$$f_n^* = f_{\mathrm{SC}}(\mathbf{x}, \mathbf{d}_n^*, \theta_{\mathrm{SC}}), \tag{4}$$

which weights the relevance of bounding box $\mathbf{d}_n^*$, we can use the Bayesian optimisation proposed in (Zhang et al. (2015)), which is an effective way to improve the detection accuracy when $f_{\mathrm{SC}}(.)$ is a computationally expensive function. The main goal of this hypothesis refinement is to improve the scale and localisation of the bounding boxes coming from (3) that can have small overlap ratios (in $[0.2, 1.0]$) with respect to the ground truth annotation. Hence, we need the scoring function defined in (4), where positive training samples are defined by an overlap≥ 0.6 and negative samples have overlap≤ 0.3. With the scoring function in (4), we can form a set $\mathcal{B}_N = \{(\mathbf{d}_n^*, f_n^*)\}_{n=1}^N$, and the goal is to find a new bounding box $\mathbf{d}_{N+1}^*$ that maximises the probability of improving the score w_{N+1}, where f is assumed to be sampled from $P(f|\mathcal{B}_N) \propto P(\mathcal{B}_N|f)P(f)$. This represents a recursive algorithm that samples a new bounding box $\mathbf{d}_{N+t}^*$ based on $\mathcal{B}_{N+t-1}$, and forms a new hypothesis set $\mathcal{B}_{N+t} = \{(\mathbf{d}_n^*, f_n)\}_{n=1}^{N+t-1} \bigcup (\mathbf{d}_{N+t}^*, f_{N+t}^*)$.

The idea behind this optimisation process is to define a prior distribution $P(f)$, defined by a Gaussian process $\mathcal{GP}(m(.), k(.,.))$, from where we can draw samples with $f \sim \mathcal{GP}(m(.), k(.,.))$ (Zhang et al. (2015)). This idea is realised with the formulation of this problem as a Gaussian regression that estimates new bounding boxes $\mathbf{d}_{N+t}^*$ given observations $\mathcal{B}_{N+t-1}$ in order to maximise the following acquisition function:

$$a_{\mathrm{EI}}(\mathbf{d}_{N+t}^*|\mathcal{B}_{N+t-1}, \theta_{EI}) = \int_{\hat{f}_N}^{\infty} (f_{N+t} - \hat{f}).P(f_{N+t}|\mathbf{d}_{N+t}^*, \mathcal{B}_{N+t-1}, \theta_{EI})df, \tag{5}$$

where $\hat{f}_N = \max_{n \in \{1,...,N\}} f_n$, θ_{EI} represents the parameters of model $a_{\mathrm{EI}}(.)$, and $P(f_{N+t}|\mathbf{d}_{N+t}^*, \mathcal{B}_{N+t-1}, \theta_{\mathrm{EI}})$ follows a Gaussian distribution (Zhang et al. (2015)). The refinement algorithm proceeds according to the steps in Algorithm 1, where non-max suppresion (NMS) is a function that takes a set of bounding boxes and clusters them based on their overlap and scores, and in-

Algorithm 1 Local search for Hypothesis Refinement

```
Require: Mammogram $\mathbf{x}$, the set of detected bounding boxes and scores $\mathcal{B}_N = \{(\mathbf{d}_n^*, f_n^*)\}_{n=1}^N$, parameters $\theta_{\text{SC}}$ for the scoring function in (4), acquisition function parameters $\theta_{\text{EI}}$ in (5), and maximum number of iterations $t_{\max}$, a threshold $f_{\text{prune}}$ to prune the bounding boxes.
$\mathcal{B}_{\text{new}} \leftarrow \text{transformations}(\mathcal{B}_N)$
for $t = 1, ..., t_{\max}$ do
    $\mathcal{B}_{\text{proposal}} = \emptyset$
    $\mathcal{B}_{\text{prune}} = \{(\mathbf{d}, f) \in \mathcal{B}_j : f \geq f_{\text{prune}}\}$
    $\mathcal{B}_{\text{nms}} = \text{NMS}(\mathcal{B}_{\text{prune}})$
    for $(\mathbf{d}_{\text{best}}, f_{\text{best}}) \in \mathcal{B}_{\text{nms}}$ do
        for $\rho \in \{0.3, 0.5, 0.7\}$ do
            $\mathcal{B}_{\text{local}} = \{(\mathbf{d}, f) \in \mathcal{B}_j : \text{IoU}(\mathbf{d}, \mathbf{d}_{\text{best}}) > \rho\}$
            $\mathbf{d}_{N+1} = \arg\max_{\mathbf{d}} a_{\text{EI}}(\mathbf{d}|\mathcal{B}_{\text{local}}, \theta_{\text{EI}})$
            $f_{N+1} = f_{\text{SC}}(\mathbf{d}_{N+1}, \mathbf{x}; \theta_{\text{SC}})$
            $\mathcal{B}_{\text{proposal}} \leftarrow \mathcal{B}_{\text{proposal}} \cup (\mathbf{d}_{N+1}, f_{N+1})$
        end for
    end for
    $\mathcal{B}_{\text{new}} \leftarrow \mathcal{B}_{\text{proposal}} \cup \mathcal{B}_{\text{new}}$
end for
$\mathcal{B}_{\text{prune}} = \{(\mathbf{d}, f) \in \mathcal{B}_{\text{new}} : f \geq f_{\text{prune}}\}$
$\mathcal{B}_{\text{ref}} = \text{NMS}(\mathcal{B}_{\text{prune}})$
```

tersection over union (IoU) measures the ratio between the intersection and the union between the two bounding boxes in the argument. In essence, Algorithm 1 runs for $t_{\max}$ steps, where we first augment the set $\mathcal{B}_N$ with the transformations(.) function that translates (in the range of $[-20, +20]$ pixels in horizontal and vertical directions, with step size 4) and scales (in the range of $[0.8, 1.2]$, with step size 0.2) the samples in $\mathcal{B}_N$ to form the set $\mathcal{B}_{\text{new}}$. Then, at each step, we first prune all candidates with low scores, and cluster the remaining ones via non-max suppression (NMS), where the assumption is that each cluster represents one particular mass candidate. For each bounding box that has been considered to be a local optimum, we consider different IoU values ($\rho \in \{0.3, 0.5, 0.7\}$) to build the local bounding box set $\mathcal{B}_{\text{local}}$ that is used in the GP to form $\mathbf{d}_{N+1}$ that is then included in the new set of proposals. This process returns the set $\mathcal{B}_{\text{ref}}$ of final mass candidates.

The estimation of the parameters θ_{SC} of the model in (5) uses the manu-

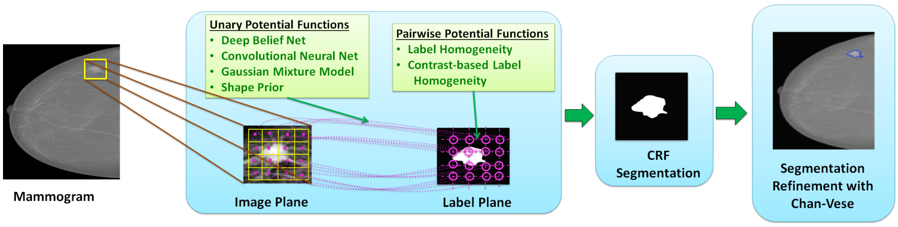


Figure 4: The proposed mass segmentation is carried out with the segmentation produced by a CRF on a low resolution image patch that is then scaled to the original image size and refined with the Chan-Vese active contour method (Chan et al. (2001)).

ally annotated bounding boxes $\mathbf{d}$ from the training data $\mathcal{D}$, which are randomly scaled and translated with positive samples comprising the bounding boxes with IoU ratio above a pre-defined threshold ρ (with respect to the manual annotation), and negative samples have IoU below that same threshold. We use the same pre-processing (contrast enhancement) (Ball and Bruce (2007)) and scaling (to an image patch of size $M \times M$) as used in Sec. 3.2.1. Finally, the model in (4) is represented by a CNN that is trained with the same samples as the ones used for training the model in (5).

4. Mass Segmentation

The mass segmentation algorithm (Dhungel et al. (2015b)) uses deep structured output learning to produce a segmentation on a low resolution input image patch. The contribution of this paper comprises a refinement step based on the Chan-Vese active contour model (Jorstad and Fua (2014)) that improves the segmentation precision in the original image resolution (see Fig. 4). Once each bounding box $\mathbf{d}_n \in \mathcal{B}_{\text{ref}}$ is estimated from the hypothesis refinement in Alg. 1, we use it to crop the image patch that is resized to a low resolution patch of size $M \times M$ with the function $\widehat{\mathbf{x}}_n = f_{\text{crop}}(\mathbf{x}, \mathbf{d}_n)$ (this function uses bi-cubic interpolation). The segmentation map is estimated in this low resolution image patch. The model used for segmenting the image is based on a Conditional Random Field (CRF), where the underlying graph $\mathcal{G}$ has nodes $\mathcal{V}$ (representing pixel grey values and labels) and edges $\mathcal{E}$ between the label nodes. The CRF model is parametrised by θ_{CRF}, where the learning minimises the following empirical loss (Nowozin and Lampert (2011)):

$$\hat{\theta}_{\text{CRF}} = \arg\min_{\theta} \sum_{i=1}^{|\mathcal{D}|} \sum_{n=1}^{|\mathcal{B}_{\text{ref}}(i)|} \ell(\widehat{\mathbf{x}}_{i,n}, \widehat{\mathbf{y}}_{i,n}, \theta), \tag{6}$$

where i indexes the training images from set $\mathcal{D}$ and n indexes the masses in the set of refined detections $\mathcal{B}_{\text{ref}}$ (with cardinality $|\mathcal{B}_{\text{ref}}|$), $\widehat{\mathbf{y}}_{n,i}$ denotes the cropped segmentation map obtained with $f_{\text{crop}}(\mathbf{y}_i, \mathbf{d}_n)$, defined above, $\ell(\widehat{\mathbf{x}}_{i,n}, \widehat{\mathbf{y}}_{i,n}, \theta)$ is a continuous and convex loss function that defines the structured output model. Our segmentation model in (Dhungel et al. (2015b)) explores CRF and SSVM formulations for solving (6), but in this paper we only consider the CRF model given its superior results. The loss function for the CRF model is described as (Dhungel et al. (2015b)):

$$\ell(\widehat{\mathbf{x}}_{i,n}, \widehat{\mathbf{y}}_{i,n}, \theta_{\text{CRF}}) = A(\widehat{\mathbf{x}}_{i,n}, \theta_{\text{CRF}}) - E(\widehat{\mathbf{x}}_{i,n}, \widehat{\mathbf{y}}_{i,n}, \theta_{\text{CRF}}), \tag{7}$$

where $A(\widehat{\mathbf{x}}_{i,n}, \theta_{\text{CRF}}) = \log \sum_{\widehat{\mathbf{y}} \in \mathbf{m} \in \{-1,+1\}^{M \times M}} \exp \{E(\widehat{\mathbf{x}}_{i,n}, \widehat{\mathbf{y}}, \theta_{\text{CRF}})\}$ is the log-partition function that ensures normalisation, and

$$\begin{aligned} E(\widehat{\mathbf{x}}_{i,n}, \widehat{\mathbf{y}}_{i,n}, \theta_{\text{CRF}}) = & \sum_{k=1}^{K} \sum_{v \in \mathcal{V}} \theta_{1,k} \psi^{(1,k)}(\widehat{\mathbf{y}}_{i,n}(v), \widehat{\mathbf{x}}_{i,n}) + \\ & \sum_{l=1}^{L} \sum_{(v,q) \in \mathcal{E}} \theta_{2,l} \psi^{(2,l)}(\widehat{\mathbf{y}}_{i,n}(v), \widehat{\mathbf{y}}_{i,n}(q), \widehat{\mathbf{x}}_{i,n}), \end{aligned} \tag{8}$$

with $\psi^{(1,k)}(.,.)$ representing one of the K unary potential functions between label and pixel nodes, $\psi^{(2,l)}(.,.,.)$ denoting one of the L binary potential functions on the edges between label nodes, and $\theta_{\text{CRF}} = [\theta_{1,1}, ..., \theta_{1,K}, \theta_{2,1}, ..., \theta_{2,L}]^{\top} \in \mathbb{R}^{K+L}$ with $\widehat{\mathbf{y}}_{i,n}(v)$ being the node v of graph $\mathcal{G}$.

4.1. Training and Inference Procedure

The solution of optimisation in (6) involves the computation of the log-partition function $A(\widehat{\mathbf{x}}_{i,n}, \theta_{\text{CRF}})$ that can be bounded from above using the tree re-weighted (TRW) belief propagation, as follows (Wainwright et al. (2003)):

$$A(\widehat{\mathbf{x}}_{i,n};\theta_{\text{CRF}}) = \max_{\mu \in \mathcal{M}} \theta_{\text{CRF}}^{T}\mu + H(\mu), \tag{9}$$

where $\mathcal{M} = \{\mu' : \exists \mathbf{w}, \mu' = \mu\}$ is the marginal polytope, $\mu = \sum_{\widehat{\mathbf{y}} \in \{-1,+1\}^{M \times M}} P(\widehat{\mathbf{y}}|\widehat{\mathbf{x}}, \theta_{\text{CRF}}) f_{\text{I}}(\widehat{\mathbf{y}})$, with $f_{\text{I}}(\widehat{\mathbf{y}})$ denoting the set of indicator functions of possible configurations of each clique and variable in the graph (Meltzer et al. (2009)), as denoted in (8), $P(\widehat{\mathbf{y}}|\widehat{\mathbf{x}}, \theta_{\text{CRF}}) = \exp\{E(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}; \theta_{\text{CRF}}) - A(\widehat{\mathbf{y}}; \theta_{\text{CRF}})\}$ indicating the conditional probability of the annotation $\widehat{\mathbf{y}}$ given the sub-image $\widehat{\mathbf{x}}$ and parameters θ_{CRF} (Assuming that this conditional probability function belongs to the exponential family) and $H(\mu) = -\sum_{\widehat{\mathbf{y}} \in \{-1,+1\}^{M \times M}} P(\widehat{\mathbf{y}}|\widehat{\mathbf{x}}; \theta_{\text{CRF}}) \log P(\widehat{\mathbf{y}}|\widehat{\mathbf{x}}, \theta_{\text{CRF}})$ is the entropy. Note that for general graphs with cycles, the marginal polytope $\mathcal{M}$ is difficult to characterise and the entropy $\mathbf{H}(\mu)$ is not tractable (Domke (2013)). TRW solves these issues by first replacing the marginal polytope with a superset $\mathcal{L} \supset \mathcal{M}$ that only accounts for the local constraints of the marginals, and then approximating the entropy calculation with an upper bound (Domke (2013)). The estimation of θ_{CRF} in (7) is achieved via gradient descent via truncated fitting (Domke (2013)), and the inference to find the label $\widehat{\mathbf{y}}^{*}$ for a sub-image $\widehat{\mathbf{x}}$ is based on TRW.

4.1.1. Potential Functions

The model in (8) can incorporate K unary and L binary potential functions. For the unary functions, we use the results from the pixel-wise segmentation produced by CNN, DBN, GMM and shape prior models. The CNN unary potential function is defined by (LeCun and Bengio (1995); Dhungel et al. (2015b))

$$\psi^{(1,1)}(\widehat{\mathbf{y}}(v), \widehat{\mathbf{x}}) = -\log P_{\text{CNNSEG}}(\widehat{\mathbf{y}}(v)|\widehat{\mathbf{x}}, \theta_{\text{CNNSEG}}), \tag{10}$$

where $P_{\text{CNNSEG}}(.)$ denotes the probability of labelling the node $v \in \mathcal{V}$ with mass or background (given the input sub-image $\widehat{\mathbf{x}}$) and θ_{CNNSEG} denotes the CNN parameters (LeCun and Bengio (1995)).

The DBN unary potential function is defined as (Hinton and Salakhutdinov

(2006); Dhungel et al. (2015b)):

$$\psi^{(1,2)}(\widehat{\mathbf{y}}(v), \widehat{\mathbf{x}}_S) = -\log P_{\text{DBNSEG},S}(\widehat{\mathbf{y}}(v)|\widehat{\mathbf{x}}_S, \theta_{\text{DBNSEG},S}), \tag{11}$$

where $\theta_{\text{DBNSEG,S}}$ represents the DBN parameters of the DBN model that receives as input an image patch of variable size centred at the node v position. The inference is based on the mean field approximation of the values in all DBN layers, followed by the computation of free energy on the top layer (Hinton and Salakhutdinov (2006)). In addition to the CNN and DBN patch-based potential functions, we also use a pixel-wise GMM unary potential function (Dhungel et al. (2015b)) defined by:

$$\psi^{(1,3)}(\widehat{\mathbf{y}}(v), \widehat{\mathbf{x}}) = -\log P_{\text{GMMSEG}}(\widehat{\mathbf{y}}(v)|\widehat{\mathbf{x}}(v), \theta_{\text{GMMSEG}}), \tag{12}$$

where $P(.)$ is computed from the GMM class dependent probability model, learned from the training set; and the shape prior unary potential function (Dhungel et al. (2015b)) is represented by

$$\psi^{(1,4)}(\widehat{\mathbf{y}}(v), \widehat{\mathbf{x}}) = -\log P(\widehat{\mathbf{y}}(v)|\theta_{\text{PRIORSEG}}), \tag{13}$$

which computes the probability that node v is part of a mass based only on the patch position (this prior is estimated from the training annotations). Finally, the pairwise potential functions between label nodes in (8) encode label and contrast dependent labelling homogeneity as $\psi^{(2,1)}(\widehat{\mathbf{y}}(v), \widehat{\mathbf{y}}(q), \widehat{\mathbf{x}})$ and $\psi^{(2,1+n)}(\widehat{\mathbf{y}}(v), \widehat{\mathbf{y}}(q), \widehat{\mathbf{x}})$ respectively (Nowozin and Lampert (2011); Domke (2013); Dhungel et al. (2015d)). The labelling homogeneity is defined by:

$$\psi^{(2,1)}(\widehat{\mathbf{y}}(v), \widehat{\mathbf{y}}(q), \widehat{\mathbf{x}}) = 1 - \delta(\widehat{\mathbf{y}}(v) - \widehat{\mathbf{y}}(q)), \tag{14}$$

where, $\delta(.)$ represents the Dirac delta function. Similarly, contrast dependent labelling homogeneity is represented by 11 pairwise potential functions and is defined by:

$$\begin{aligned}\psi^{(2,1+n)}(\widehat{\mathbf{y}}(v),\widehat{\mathbf{y}}(q),\widehat{\mathbf{x}}) &= (1-\delta(\widehat{\mathbf{y}}(v)-\widehat{\mathbf{y}}(q))\delta(||\lfloor\widehat{\mathbf{x}}(v)\rfloor_{\tau_n}-\lfloor\widehat{\mathbf{x}}(q)\rfloor_{\tau_n}||_2)),\\ \lfloor\widehat{\mathbf{x}}(v)\rfloor_{\tau_n} &= \begin{cases}\widehat{\mathbf{x}}(v) \text{ if } \widehat{\mathbf{x}}(v)\geq\tau_n\\ 0, \text{ otherwise,}\end{cases}\end{aligned} \tag{15}$$

where $\widehat{\mathbf{x}}(v), \widehat{\mathbf{x}}(q)$ represents the value of the pixel at grid location v, q, and $\tau_n \in \{\tau_1, \tau_2, ..., \tau_{10}\}$ is a set of ten thresholds (Domke (2013); Dhungel et al. (2015d)).

4.2. Segmentation Refinement

We map the segmentation $\widehat{\mathbf{y}}^*$, obtained from the inference described in Sec. 4.1, from the $M \times M$ lattice to the original image size, using the bounding box $\mathbf{d}_n \in \mathcal{B}_{\text{ref}}$ with the function $\widetilde{\mathbf{y}}^*_n = f_{\text{restore}}(\widehat{\mathbf{y}}^*, \mathbf{d}_n)$ that uses nearest neighbour interpolation. The issue here is that the resulting segmentation $\widetilde{\mathbf{y}}^*_n$ is quite coarse and needs to be refined, and our solution involves the use of the Chan-Vese active contour (Chan et al. (2001)) with $\widetilde{\mathbf{y}}^*_n$. The active contour function $\phi(.)$ to represent the segmentation is the signed distance function and $\widetilde{\mathbf{y}}^*_n$ is used to initialise this function with $\phi_0 = f_\phi(\widetilde{\mathbf{y}}^*)$, where the energy functional to be minimised is defined by (Chan et al. (2001)):

$$\begin{aligned}E_{\text{CV}}(\phi, \widetilde{\mathbf{y}}^*, \mathbf{x}) = \gamma \int_\Omega |(\mathbf{x}-c_2)|^2(1-H(\phi)dx+\\ \lambda \int_\Omega |(\mathbf{x}-c_1)|^2 H(\phi)dx + \mu \int_\Omega \delta(\phi)|\bigtriangledown \phi|dx,\end{aligned} \tag{16}$$

where $H(.)$ is the heaviside step function, μ, λ, γ are tunable parameters, c_1, c_2 are the average of the image $\mathbf{x}$ in the regions where $\phi(.) >= 0$ and $\phi(.) < 0$ (respectively), and $\delta(.)$ is the Dirac delta function. The minimisation of the energy in (16) is solved by finding the steady state solution of the gradient flow equation $\frac{\partial\phi}{\partial t} = -\frac{\partial E}{\partial \phi}$, where $\frac{\partial E}{\partial \phi}$ is the Gâteaux derivative of the functional $E(.)$ (Chan et al. (2001)). The final segmentation is produced by $\mathbf{y}^*_n = \phi \geq 0$. The full segmentation algorithm is displayed in Algorithm. 2, and depicted in

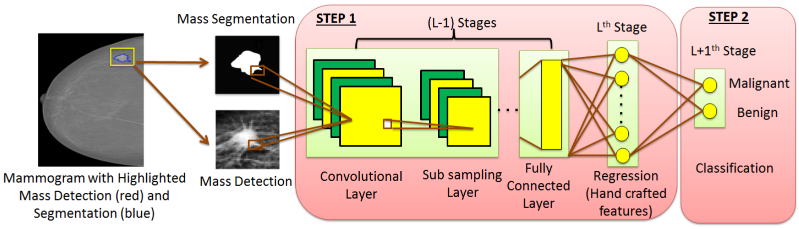


Figure 5: The proposed classification methodology consists of two steps: 1) pre-training of the CNN for regressing the values of hand-crafted features, and 2) fine-tuning the pre-trained CNN model for the mass classification problem.

Fig. 4.

Algorithm 2 Mass Segmentation with Refinement

```
Require: Mammogram $\mathbf{x}$, refined bounding box $\mathbf{d}_n \in \mathcal{B}_N$, sub-image size $M_{\text{sub}}$,
    number of iterations $t_{\max}$ for the Chan-Vese optimisation, the unary and
    pairwise model parameters $\theta_{\text{CNNSEG}}$, $\theta_{\text{DBNSEG}}$, $\theta_{\text{GMMSEG}}$, $\theta_{\text{PRIORSEG}}$, and
    structured output model $\theta_{\text{CRF}}$
  Extract sub-image $\widehat{\mathbf{x}} = f_{\text{s}}(\mathbf{d}_n, \mathbf{x}, M_{\text{sub}})$
  Constrast enhance sub-image $\widehat{\mathbf{x}}$ (Ball and Bruce (2007))
  Compute unary potential function results $\psi^{(1,k)}$ for $k \in \{1, ..., 4\}$ using (10)-
    (13)
  Compute pairwise potentials $\psi^{(2,l)}$ for $k \in \{1, 2\}$ using (Meltzer et al. (2009))

  Infer segmentation label $\widehat{\mathbf{y}}^*$ using TRW (Wainwright et al. (2003); Dhungel
    et al. (2015b))
  Map $\widehat{\mathbf{y}}^*$ to $\widetilde{\mathbf{y}}^* = f_{\text{restore}}(\widehat{\mathbf{y}}^*, \mathbf{d}_n)$
  Compute initial distance function $\phi_0 = f_\phi(\widetilde{\mathbf{y}}^*)$
  Estimate $\phi_{t_{\max}}$ using Chan-Vese minimization (Chan et al. (2001))
  Infer final segmentation $\mathbf{y}_n^* = \phi_{t_{\max}} \geq 0$
```

5. Mass Classification

The main idea explored in the implementation of the mass classification system is to leverage the functionality of previously proposed hand-crafted features (Varela et al. (2006)) in the training of the CNN model (LeCun and Bengio (1995); Krizhevsky et al. (2012)), particularly considering that such features have been shown to be effective for tumour classification. Specifically, the CNN mass classification model is trained in two stages. The first stage pre-trains the CNN model to work as a regressor from the input image patch and respective segmentation against the values of a large set of hand-crafted features as per Sec. 3.2.1. The second stage fine-tunes the pre-trained CNN model to improve the accuracy of breast mass classification.

The hand-crafted features are extracted from a mammogram $\mathbf{x}$, bounding box $\mathbf{d}$ and segmentation map $\mathbf{y}$ as follows:

$$\mathbf{z} = f_{\text{hcf}}(\mathbf{x}, \mathbf{d}, \mathbf{y}), \tag{17}$$

where $\mathbf{z} \in \mathbb{R}^H$ denotes the vector containing the values of the hand-crafted features, consisting of morphological and texture features (Varela et al. (2006)). The morphological features are computed using the segmentation mask $\mathbf{y}$, and the texture features are computed from the image patch contained by the bounding box $\mathbf{d}$ as in Sec. 3.2.1. In order to pre-train the CNN model with the features $\mathbf{z}$, we build a model with $L-2$ stages of convolutional plus non-linear activation and max pooling, followed by a fully connected layer with H nodes, which is the same number of features as in $\mathbf{z}$ in (17). This regressor is defined by

$$\mathbf{z}^* = f_{\text{CNNRG}}(\mathbf{x}, \mathbf{d}, \mathbf{y}, \theta_{\text{CNNRG}}), \tag{18}$$

where $f_{\text{CNNRG}}(.)$ represents the CNN model that outputs the estimated hand-crafted feature vector $\mathbf{z} \in \mathbb{R}^H$, where the loss function used to train such model is denoted by $\ell(\theta_{\text{CNNRG}}) = \sum_{i=1}^{|\mathcal{D}|} \sum_{j}^{|\mathcal{A}_i|} \|\mathbf{z}_{i,j} - \mathbf{z}^*_{i,j}\|_2$, with i indexing the training images, j indexing the masses in each training image, $\mathbf{z}_{i,j}$ denotes the vector of hand-crafted features from mass j and image i, and $\mathbf{z}^*_{i,j}$ is the output from (18) - see step 1 in Fig. 5. The mass classification model takes the CNN from (18) and adds another fully connected layer (i.e., the $L + 1^{st}$ layer) with softmax activation, which is trained with cross enropy loss minimisation - see step 2 in Fig. 5.

6. Experimental Methodology

We evaluate the performance of our detection, segmentation and classification methodologies on the publicly available INbreast dataset (Moreira et al. (2012)), containing 115 cases and 410 images, out of which 116 images have benign or malignant masses and the remaining ones do not contain any masses. For the experiments, the 115 cases of the dataset are randomly divided into 60% for training, 20% for validation and 20% for testing, which allows us to run a five-fold cross validation. All experiments are carried out on a standard computer with the following specification: Intel(R) Core(TM) i5-2500k 3.30GHz CPU with 8GB RAM and graphics card NVIDIA GeForce GTX 460 SE 4045 MB.

6.1. Detection Experimental Setup

For the detection experiment, we use the average precision curve, which is a function of true positive rate against the Intersection over Union (IoU), and free response operating characteristic (FROC) curve that is a function of true positive rate (TPR) with respect to false positive detections per image (FPI). For the mass ROI detection problem in Sec. 3.2.1, the mass is considered to be detected if the IoU between the bounding box of the candidate region and ground truth is greater than or equal to 0.2 (Kozegar et al. (2013); Beller et al. (2005); te Brake et al. (2000); Campanini et al. (2004); Eltonsy et al. (2007); Sampat et al. (2008); Bellotti et al. (2006); Wei et al. (2005)). The model selection for the DBN, R-CNN and RF models in mass ROI detection (Sec. 3.2.1) is performed with the training and validation sets. The network structure for the m-DBN in Sec. 3.2.1 has two layers containing 200 and 500 nodes and the input patch has a fixed size of 7×7 (i.e., $S = 7$) for all resolutions of the input image, where the coarsest resolution is represented by an image of size 80×80 (pixels), the next finer resolutions have images of sizes $120 \times 120, 160 \times 160$ and 264×264. We use the LeNet network structure (LeCun and Bengio (1995)) for both CNN models in the cascade of R-CNN models in Sec. 3.2.1, where the input image has a fixed size of 40×40 pixels (i.e., $M = 40$). The LeNet network structure has 20 filters of size 5×5 followed by a max pooling layer that sub-samples the input by a factor of two, then the second convolutional stage has 50 filters of size 5×5 and a max-pooling layer that again sub-samples the input by two, the convolutional stage three has 500 filters of size 4×4 followed by a rectified linear unit (ReLU) activation function (Nair and Hinton (2010)), the fourth convolutional stage has 500 filters with size 4×4 followed by another ReLU unit, and stage five is a fully connected layer with 2 nodes. For the R-CNN models, we artificially augment the number of positive training samples from the mass ROI detection stage using geometric transformations such as translation and rotation around the positive candidates. The augmented dataset contains 10 times the initial number of positive samples, but the original number of negative samples. The samples are considered positive if the respective bounding boxes have IoU ≥ 0.2, otherwise they are regarded as negative. The RF classifier is trained without data augmentation. The operating point for the cascaded module in mass ROI detection is fixed by setting a threshold on classifiers scores using the training and validation set which ensures that TPR $>= 0.9$ while gradually reducing the FPI in each stage of the cascade (see Fig. 3). The parameters for the RF classifiers are estimated with the validation set of each one of the five folds of the N-fold cross validation with search range from [1,1000]. On average, the first cascade stage of RF has 37 trees, with each tree containing 27 leaves, whereas the second cascade stage has 56 trees, each containing 17 leaves. The definition of positive and negative samples is the same as above for the R-CNN models, but we do not use the augmented training data.

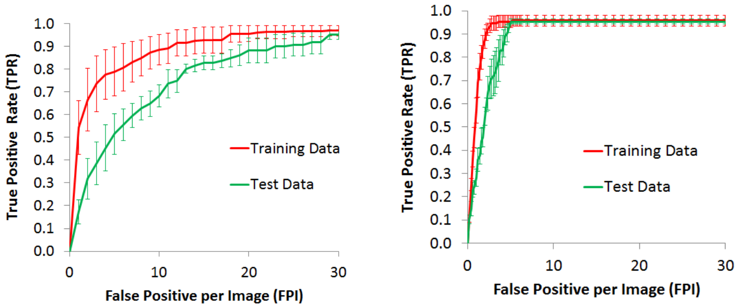
For the hypothesis refinement, we use a separate CNN model represented by θ_{SC} defined in (4), which has the LeNet network structure (LeCun and Bengio (1995)). This new classifier in (4) is important because the RF model above has a relatively low precision in terms of the detection of the position and scale of the mass, where a positive sample is defined by IoU≥ 0.3. This new CNN classifier defines a positive sample by IoU≥ 0.6 and a negative sample by IoU< 0.6. These samples are obtained by augmenting the ground truth bounding box (translation and scale) using training data followed by cropping, re-sizing with bi-cubic interpolation to 40×40 and contrast enhancement (Ball and Bruce (2007)).

6.2. Segmentation Experimental Setup

The model selection for the DBN (θ_{DBNSEG}) and CNN (θ_{CNNSEG}) unary potential functions in Algorithm. 2 is performed via cross validation using the training and validation sets. The DBN model has two layers with 200 and 500 nodes, which are trained with image patch sizes of 3×3, 5×5, and 7×7. The CNN model has two convolutional stages with 12 filters of sizes 5×5 that are followed by ReLU activation and max-pooling that reduces the input size by a factor of two. The final stage of the CNN model has a fully connected layer containing 588 nodes and an output layer of 40×40 (i.e., the same size as the input). Finally, the parameter values for the Chan-Vese model in (16) are also estimated via cross validation, producing the following values: $\mu = 0.2, \lambda = 1, \gamma = 1$ and number of iterations $t = 10$.

6.3. Classification Experimental Setup

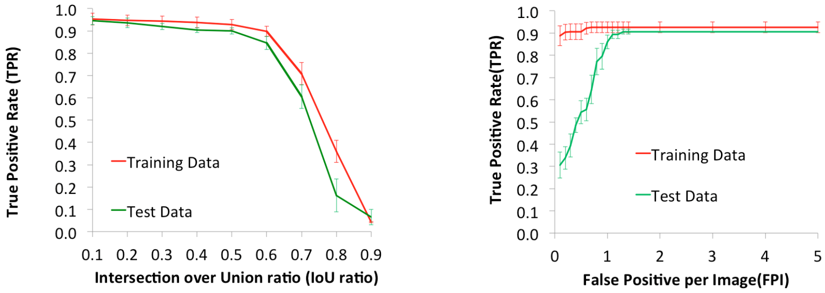
We explore both a manual and fully automated setup for classification where manual set-up uses the manual annotations for the ROI and segmentation mask. We use the refined ROI bounding box obtained from Algorithm 1 (same used for the mass segmentation) and segmentation mask from Algorithm 2 for the fully automated set-up. From the ROI bounding box and segmentation mask, we extract 781 hand-crafted features, as described in Sec. 3.2.1, for pre-training the CNN model. The CNN model that is pre-trained with these features has the first convolutional stage with 20 filters of size 5×5 followed by a max pooling layer that sub-samples the input by factor of two, then the second convolutional stage has 50 filters of size 5×5 and a max-pooling layer that again sub-samples the input by two, the convolutional stage three has 100 filters of size 4×4 followed by a rectified linear unit (ReLU) activation function (Nair and Hinton (2010)), the fourth convolutional stage has 781 filters with the size 4×4 followed



(a) FROC - Cascade of R-CNN (b) FROC - Cascade of RF

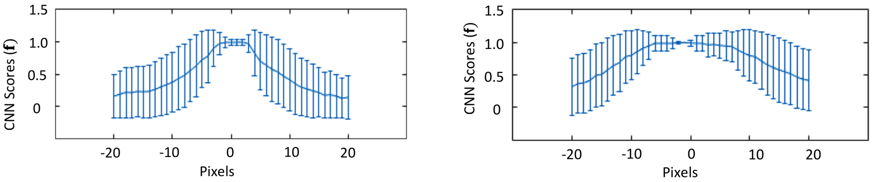
Figure 6: FROC curve for cascade of R-CNN and RF (Dhungel et al. (2015a)) during the ROI detection, assuming that a successful detection has IoU of at least 0.2 (Kozegar et al. (2013); Beller et al. (2005); te Brake et al. (2000); Campanini et al. (2004); Eltonsy et al. (2007); Sampat et al. (2008); Bellotti et al. (2006); Wei et al. (2005)).

by another ReLU unit, and stage five is a fully connected layer with 781 nodes (i.e., the same size as the hand-crafted features). The CNN model used for the fine-tuning process uses the pre-trained model, where a softmax layer containing two nodes (representing the benign versus malignant classification) is added, and the fully-connected layers are trained with drop-out of 0.3 (Srivastava et al. (2014)). In order to regularise the CNN, we artificially augment by 10 times the training data using geometric transformations (rotation, translation and scale) in the vicinity of the ground truth data. Note that for comparison purposes, we also train a CNN model without the pre-training step to show its influence in the classification accuracy. Moreover, using the hand-crafted features, we train an RF classifier (Breiman (2001)), where model selection is performed using the validation set of each cross validation fold. We also train another RF classifier using the 781 features from the second to last fully-connected layer of the fine-tuned CNN model. The parameters for the RF classifiers are estimated with the validation set of each one of the five folds of the N-fold cross validation where on average, the RFs have 8 trees (search range in [1,1000]), each with 6 leaves (search range in [1,1000]).



(a) Average precision for detection (b) FROC - Mass hypothesis refinement

Figure 7: Performance measures of our proposed mass refinement algorithm: a) True positive rate of hypothesis refinement as a function of the the minimum IoU ratio, and b) FROC curve of the hypothesis refinement at IoU ≥ 0.5.



(a) Horizontal Translation (b) Vertical Translation

Figure 8: Plot of the CNN classifier in (5) as a function of the annotated bounding box horizontal (a) and vertical (b) translation.

7. Experimental Results

Fig. 6-(a-b) shows the FROC curve as a performance measure for the cascade stages in the ROI detection module. The final mass ROI detection module, consisting of the RF in Sec. 3.2.1 produces a TPR of 0.95 ± 0.02 at a FPI $= 5$ for the testing data and TPR of 0.95 ± 0.02 at FPI $= 3$ for training data with an IoU ≥ 0.2 (see FROC curve in Fig. 6-(b)). Figure 7-(a) shows the TPR as a function of different minimum levels of IoU for the hypothesis refinement in Algorithm. 1, where it can be noted that for values where IoU ≤ 0.5, TPR remains stable and above 0.9 and starts to fall with IoU > 0.5 for both training and testing. Therefore, we choose an IoU $= 0.5$ based on the training result as an optimal point for measuring the performance of our mass detection algorithm

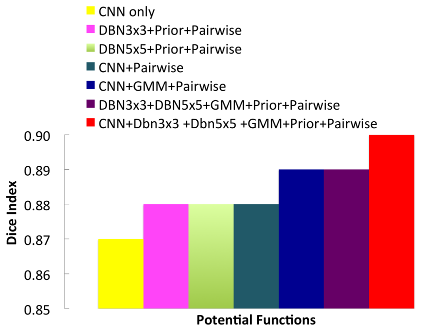


Figure 9: Effect of adding different potential functions into our CRF model (Dhungel et al. (2015b)) on the testing set of INbreast taking a manually detected ROI breast mass.

Table 1: Results of our fully automated segmentation algorithm on the INbreast dataset.

Segmentation Methodology	**Input Size**	**Dice Index (Training Data)**	**Dice Index (Test Data)**
CRF model with active contour refinement	Original image resolution	0.85 ± 0.01	0.85 ± 0.02
CRF model	40x40	0.87 ± 0.02	0.84 ± 0.02
CRF model with nearest neighbor interpolation	Original image resolution	0.82 ± 0.02	0.80 ± 0.01
Active contour model	Original image resolution	0.82 ± 0.01	0.82 ± 0.03

with the hypothesis refinement described in Sec. 3.3. From the FROC curve in Fig. 7-(b), the mass detection algorithm with hypothesis refinement produces the best result of TPR $= 0.93 \pm 0.05$ at FPI $= 0.8$ on the training data and a TPR $= 0.90 \pm 0.02$ at a FPI $= 1.3$ on the testing data with an IoU ≥ 0.5. We also found that our automated mass ROI detection and refinement system produces a pixel wise TPR of 0.99 ± 0.01 for training and a TPR of 0.97 ± 0.02 for the testing data. Fig. 8-(a) and Fig. 8-(b) show the result of the scoring function, as a function of horizontal and vertical translation of the ground truth, in the hypothesis refinement described in Sec. 3.3. The two graphs in Fig. 8 show that the scoring function has high accuracy and precision when a small translation (< 5 pixels) is applied, and both measures tend to decrease with larger translations (> 5 pixels).

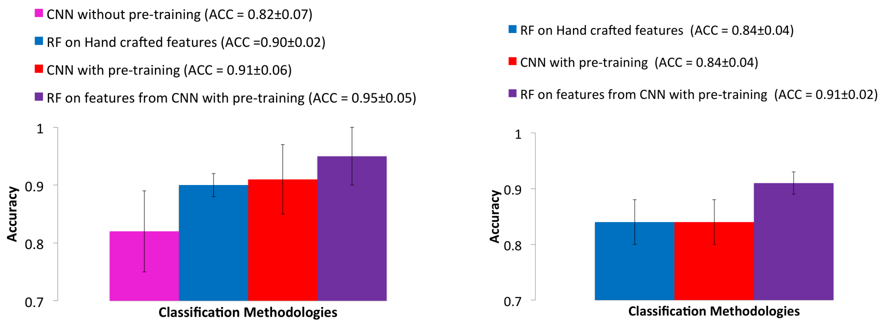
The performance of the proposed segmentation algorithm is shown in Tab. 1

Table 2: Comparison between our proposed segmentation algorithm and the state-of-the-art methods on test sets.

Methodologies	Setup	Dataset	Rep.	Dice Index
Proposed CRF model with active contour refinement	Fully automated	INbreast	yes	0.85 ± 0.02
te Brake et al. (2000)	Fully automated	Dutch screening program	no	0.82
Our previous CRF model w/o refinement (Dhungel et al. (2015b))	Semi-automated	INbreast	yes	0.90 ±0.02
Cardoso et al. (2015)	Semi-automated	INbreast	yes	0.88

in terms of the Dice index for training and testing data from the automatically detected and refined ROIs from Algorithm. 1. The segmentation was carried out using the combination of several potential functions ($\text{CNN}+\text{DBN}3\times3 + \text{DBN}5\times5 + \text{GMM} + \text{Prior} + \text{Pairwise}$) for the CRF segmentation at resolution of 40×40 (Dhungel et al. (2015b)). We also show the result in terms of Dice index for combining different potential functions to our CRF model for the segmentation of manually detected ROIs in Fig. 9 (Dhungel et al. (2015b)). The resulting segmentation in a 40×40 binary image is resized to its original bounding box size using bicubic-interpolation and then refined using Chan-Vese's active contour model (Chan et al. (2001)), as described in Sec. 4.2. For comparison, we show the Dice index of the segmentation when the segmentation map is scaled up to the original image resolution using nearest neighbour interpolation. Also for comparison, we show the result from Chan-Vese's active contour (Chan et al. (2001)) with a general initialisation with an ellipse centred and scaled according to the position and size of the bounding box. This initial ellipse shape is obtained by fitting an ellipse to all aligned training annotations. Table 2 shows a comparison between our proposed segmentation method and the current state of the art in field, where the column represented by "Rep." indicates public availability of datasets to reproduce the result and "Setup" indicates whether the mass ROI detection is performed in a fully automated way, or semi-automated manner (i.e. with a manual mass detection).

For the classification problem we compare the performance of different versions of the proposed model in order to assess the role of each stage. Figures 10-



(a) Manual setup

(b) fully automated setup

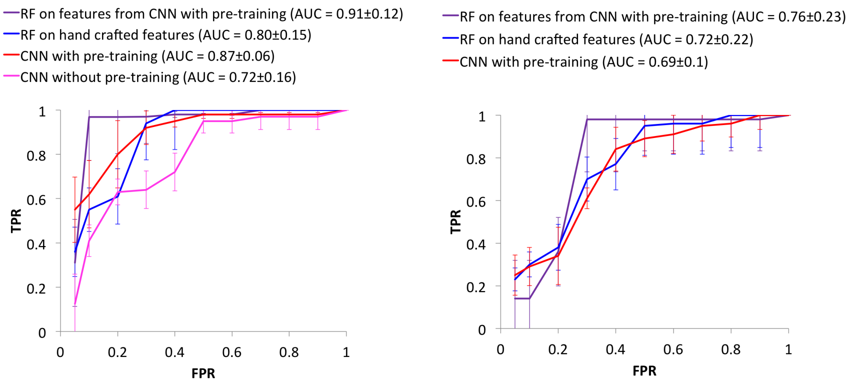
Figure 10: Accuracy of various classifier on features extracted using methodologies described in this paper in manual and automated system for test data.

Table 3: Comparison between our classification methodology and state-of-the-art methods on test sets.

Methodology	Dataset	Setup	ACC	AUC
Proposed CNN with pre-training	INbreast	Manual	0.91 ± 0.06	0.87 ± 0.06
Proposed RF on CNN with pre-training	INbreast	Manual	0.95 ± 0.05	0.91 ± 0.12
Proposed CNN with pre-training	INbreast	Fully automated	0.84 ± 0.04	0.69 ± 0.10
Proposed RF on CNN with pre-training	INbreast	Fully automated	0.91 ± 0.02	0.76 ± 0.23
Domingues et al. (2012)	INbreast	Manual	0.89	NA
Varela et al. (2006)	DDSM	Semi-automated	0.81	0.76
Ball and Bruce (2007)	DDSM	Semi-automated	0.87	0.97
Shi et al. (2007)	Uni. of Michigan	Semi-automated	0.83 ± 0.02	0.85 ± 0.02

(a-b) displays the classification accuracy for both manual and automated setups, from which it is apparent that the RF on the features from the CNN model with pre-training produces the best results on the testing set with an accuracy (ACC) of 0.95 ± 0.05 on manual and 0.91 ± 0.02 on the fully automated setup. In addition, we compare the results between the various models in terms of area under the ROC curve (AUC) in Figures 11-(a-b), which also shows that RF on the CNN features with pre-training produces the best overall AUC value of 0.91 ± 0.12 for manual and 0.76 ± 0.23 for fully automated setup. We also compare our classification method with other state-of-the-art methods in Tab. 3 in terms of classification accuracy and AUC where applicable.

The total running time for our fully automated system is 41 seconds per



(a) Manual setup

(b) fully automated setup

Figure 11: ROC curve of various classifier on features extracted using methodologies described in this paper in manual and automated system for test data.

image, divided into 39 seconds for mass detection, 0.2 seconds for the mass segmentation and 0.8 seconds for mass classification. We show some visual results in Fig. 12 for the fully automated detection and segmentation results and in Fig. 13 for the fully automated detection, segmentation and classification system.

8. Discussion

The results from the Fig. 7-(a-c) show the importance of hypothesis refinement stage of the segmentation algorithm in Algorithm. 1. This improves the localisation precision of the bounding box, and consequently increases the IoU ratio with respect to the ground truth annotation from 0.2 to 0.5 while keeping TPR over 0.9 and FPI around one. The other important observation is that our proposed mass detection algorithm retains most of ground truth pixels in training (99%) as well as testing (97%). The FROC curves in Fig. 6 show the benefit of the proposed cascade classifier. The TPR from the second cascade stage of R-CNN saturates when FPI is around 30 without making any further improvement. We also noticed that it is important to have two stages of R-CNN because a single R-CNN module is not enough to reduce the FPI to around 30

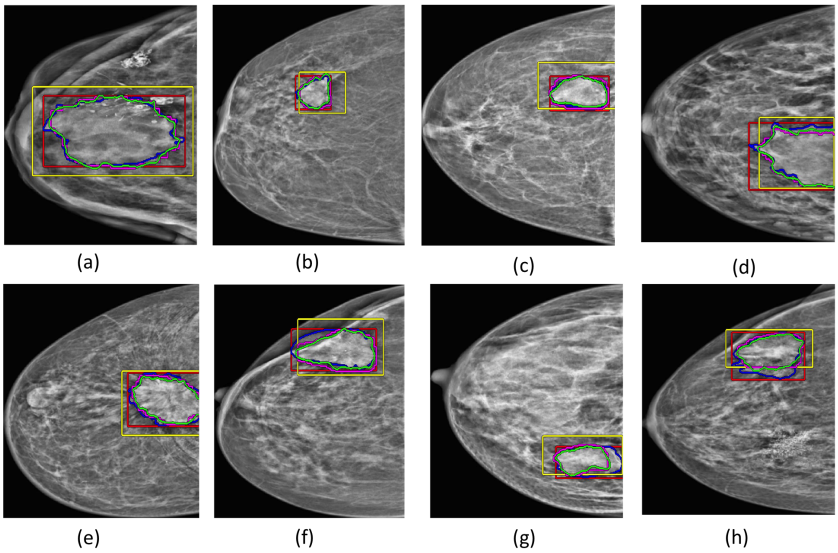


Figure 12: Examples of the fully automated mass detection and segmentation with refinement. The contour with the blue line represents the ground truth annotation, red line denotes the manual ROI, yellow is the detected and refined ROI from our methodology, magenta is the segmentation from the CRF model with nearest neighbor interpolation, and green is the segmentation refined by the active contour model.

(at a TPR $\geq$ 0.95). We also found that in order to achieve the best performance for the hypothesis refinement module, it is important to reduce the FPI to around five whilst keeping the TPR above 0.9. In this sense, the proposed cascade with two RF stages plays an important role as a single stage of RF was not able to achieve acceptable results.

The segmentation result in Fig. 9 (Dhungel et al. (2015b)) on manual setup shows that the combination of all the potential functions (CNN + DBN3x3 + DBN5x5 + GMM + prior + pairwise) is crucial for producing state-of-the-art results. Therefore, we use all these potential functions in our CRF segmentation model for the fully automated setting. The segmentation results in Table. 1 show that the proposed model with active contour refinement produces better results (Dice Index $= 0.85 \pm 0.02$) on the testing set compared with nearest neighbour interpolation from the 40×40 CRF result to the original image resolution (Dice

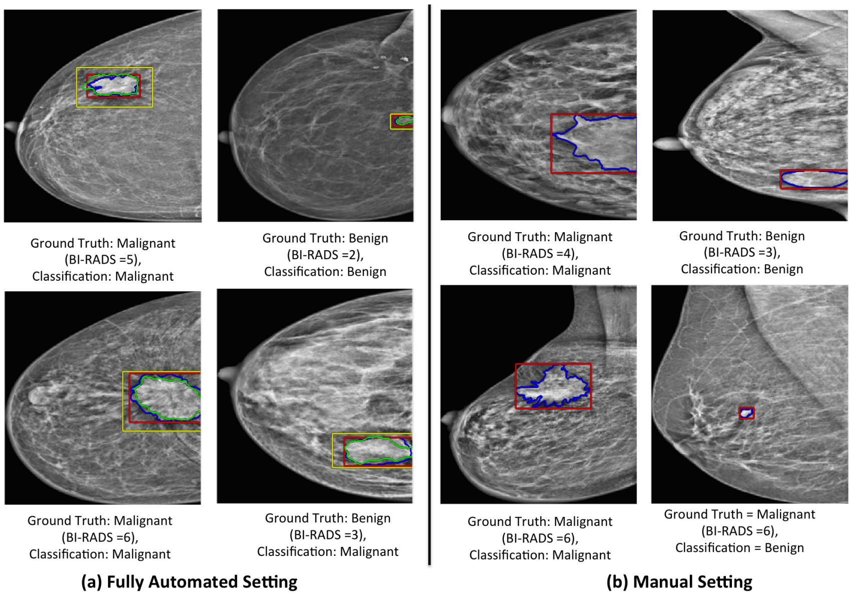


Figure 13: Examples of mass classification based on the RF model on features from CNN with pre-training using the fully automated setup and manual setup. Red contours denote manual detection and blue denotes the manual segmentation whereas yellow contours represent the automated detection and green is the automated segmentation. Ground truth and automated classification results are shown below each image.

Index $= 0.82 \pm 0.02$) and the active contour model with a fixed initialisation computed from the mean shape of the training set (Dice Index $= 0.82 \pm 0.01$). It is also important to notice that the proposed segmentation refinement produces slightly better results on test data when compared with the CRF model on the 40×40 resolution. We also notice that the number of iterations needed for the active contour model to converge using segmentation from the proposed CRF model is smaller (10 iterations) than the number of iterations needed when using the mean shape from training set (100 iterations). The comparison with the current state-of-the-art systems for segmentation in Table. 2 shows that our methodology produces the best result when using automatically generated mass ROIs (Dice Index $= 0.85 \pm 0.02$ vs 0.82 (te Brake et al. (2000))) as well in manually selected ROIs (Dice Index $= 0.90$ vs 0.88 (Cardoso et al. (2015))).

For the mass classification problem, the results in Figures 10 and 11 show that RF on features from the CNN model with the pre-training and CNN with pre-training are better than the results using RF on hand-crafted features and CNN without pre-training. Figures 10 and 11 also show that the RF classifier performs better than the CNN classifier in both fully automated and manual setups. Here, we did not show the classification results of CNN without pre-training for the fully automated system because of its poor performance on manual setup. The Wilcoxon paired signed-rank for classification accuracy on test set between the RF on CNN features with pre-training and the RF on hand-crafted features indicates statistically significant results (at 5% level), with a p-value of 0.02. Another important observation from the Table. 3 is that both the training accuracy (ACC = 0.94±0.06) and testing accuracy (ACC = 0.95±0.05) on manual setup correlates well with each other implying good generalisation of RF on CNN features with pre-training. From the Fig.11 (a-b), we see that there is an increase in FPR and decrease in the AUC value in fully automated system compared to manual setup which is expected due to increase in number of FPI in fully automated setup. Table. 3 shows that our methodology produces comparable or better results in terms of classification accuracy in manual, semi-automated and fully automated setups. The visual results in Fig. 13-(a) shows classification results using fully automated set-up and Fig. 13-(b) shows the results from the manual set-up. The visual results for the fully automated set-up has quite an accurate automatically generated ROI and segmentation using our technique. Finally, the classification results on test set, using manual set-up, display a mean sensitivity of 0.97 and mean specificity of 0.90, while the fully automated set-up produces a mean sensitivity of 0.98 and mean specificity of 0.70, which shows that our proposed CAD system is robust to false positives and false negatives.

9. Conclusion

In this paper, we describe a complete and fully automated CAD system for detection, segmentation and classification of masses from mammograms. Our mass detection method consists of a cascade of deep learning and random forest models for the generation of mass candidates and reduction of false positives, followed by hypothesis (detection) refinement. Segmentation is then carried out with the sub-image extracted from the detected masses, which is refined by typical active contour models to provide more accurate delineation in higher resolution images. The refined hypothesis and respective refined segmentation mask are then used in a two-step training process for mass classification using a CNN model, where pre-training is done in the first step in order to approximate the values of hand-crafted features, and then it is fine-tuned for the breast mass classification problem. In general, our fully automated mass detection, segmentation and classification system produces promising results and can be used as baseline result. We also believe that our current methodology can be incorporated in the clinical set-up as a second reader for radiologists. In future, we would like to build a end-to-end system capable of detection, segmentation and classification in a single integrated module similar to that of Fast R-CNN (Girshick (2015)) which has produced state-of-the-art result recently in the field of object detection.

References

AIHW, 2012. Breast cancer in australia: an overview. Cancer series no. 71. Cat. no. CAN 67, Canberra: AIHW .

Ball, J.E., Bruce, L.M., 2007. Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation, in: Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, IEEE. pp. 4973–4978.

Beller, M., Stotzka, R., Müller, T.O., Gemmeke, H., 2005. An example-based system to support the segmentation of stellate lesions, in: Bildverarbeitung für die Medizin 2005. Springer, pp. 475–479.

Bellotti, R., De Carlo, F., Tangaro, S., Gargano, G., Maggipinto, G., Castellano, M., Massafra, R., Cascio, D., Fauci, F., Magro, R., et al., 2006. A completely automated cad system for mass detection in a large mammographic database. Medical physics 33, 3066–3075.

te Brake, G.M., Karssemeijer, N., Hendriks, J.H., 2000. An automatic method to discriminate malignant masses from normal tissue in digital mammograms. Physics in Medicine and Biology 45, 2843.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Campanini, R., Dongiovanni, D., Iampieri, E., Lanconelli, N., Masotti, M., Palermo, G., Riccardi, A., Roffilli, M., 2004. A novel featureless approach to mass detection in digital mammograms based on support vector machines. Physics in Medicine and Biology 49, 961.

Cardoso, J.S., Domingues, I., Oliveira, H.P., 2015. Closed shortest path in the original coordinates with an application to breast cancer. International Journal of Pattern Recognition and Artificial Intelligence 29, 1555002.

Carneiro, G., Nascimento, J., Bradley, A.P., 2015. Unregistered multiview mammogram analysis with pre-trained deep learning models, in: Medical Image Computing and Computer-Assisted InterventionMICCAI 2015. Springer, pp. 652–660.

Chan, T.F., Vese, L., et al., 2001. Active contours without edges. Image processing, IEEE transactions on 10, 266–277.

Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, pp. 411–418.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273–297.

Dhungel, N., Carneiro, G., Bradley, A., 2015a. Automated mass detection in mammograms using cascaded deep learning and random forests, in: Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on, pp. 1–8. doi:10.1109/DICTA.2015.7371234.

Dhungel, N., Carneiro, G., Bradley, A.P., 2015b. Deep learning and structured prediction for the segmentation of mass in mammograms, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Springer, pp. 605–612.

Dhungel, N., Carneiro, G., Bradley, A.P., 2015c. Deep structured learning for mass segmentation from mammograms, in: Image Processing (ICIP), 2015 IEEE International Conference on, pp. 2950–2954. doi:10.1109/ICIP.2015.7351343.

Dhungel, N., Carneiro, G., Bradley, A.P., 2015d. Tree re-weighted belief propagation using deep learning potentials for mass segmentation from mammograms, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), pp. 760–763. doi:10.1109/ISBI.2015.7163983.

Dhungel, N., Carneiro, G., Bradley, A.P., 2016. The automated learning of deep features for breast mass classification from mammograms, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016. Springer, p. Accepted for publication.

Domingues, I., Sales, E., Cardoso, J., Pereira, W., 2012. Inbreast-database masses characterization. XXIII CBEB .

Domke, J., 2013. Learning graphical model parameters with approximate marginal inference. arXiv preprint arXiv:1301.3193 .

Dromain, C., Boyer, B., Ferre, R., Canale, S., Delaloge, S., Balleyguier, C., 2013. Computed-aided diagnosis (cad) in the detection of breast cancer. European journal of radiology 82, 417–423.

Elmore, J.G., Jackson, S.L., Abraham, L., et al., 2009. Variability in interpretive performance at screening mammography and radiologists characteristics associated with accuracy1. Radiology 253, 641–651.

Eltonsy, N.H., Tourassi, G.D., Elmaghraby, A.S., 2007. A concentric morphology model for the detection of masses in mammography. Medical Imaging, IEEE Transactions on 26, 880–889.

Farabet, C., Couprie, C., Najman, L., LeCun, Y., 2013. Learning hierarchical features for scene labeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35, 1915–1929.

Fenton, J.J., Taplin, S.H., Carney, P.A., Abraham, L., Sickles, E.A., D'Orsi, C., Berns, E.A., Cutter, G., Hendrick, R.E., Barlow, W.E., et al., 2007. Influence of computer-aided detection on performance of screening mammography. New England Journal of Medicine 356, 1399–1409.

Giger, M.L., Pritzker, A., 2014. Medical imaging and computers in the diagnosis of breast cancer, in: SPIE Optical Engineering+ Applications, International Society for Optics and Photonics. pp. 918908–918908.

Girshick, R., 2015. Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE. pp. 580–587.

Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural computation 18, 1527–1554.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T., Thun, M.J., 2008. Cancer statistics, 2008. CA: a cancer journal for clinicians 58, 71–96.

Jorstad, A., Fua, P., 2014. Refining mitochondria segmentation in electron microscopy imagery with active surfaces, in: Computer Vision-ECCV 2014 Workshops, Springer. pp. 367–379.

Kozegar, E., Soryani, M., Minaei, B., Domingues, I., et al., 2013. Assessment of a novel mass detection algorithm in mammograms. Journal of cancer research and therapeutics 9, 592.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

LeCun, Y., Bengio, Y., 1995. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361.

Meltzer, T., Globerson, A., Weiss, Y., 2009. Convergent message passing algorithms: a unifying view, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press. pp. 393–401.

Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S., 2012. Inbreast: toward a full-field digital mammographic database. Academic Radiology 19, 236–248.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814.

Ngo, T.A., Carneiro, G., 2014. Fully automated non-rigid segmentation with distance regularized level set evolution initialized and constrained by deep-structured inference, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE. pp. 3118–3125.

Nowozin, S., Lampert, C., 2011. Structured learning and prediction in computer vision. Foundations and Trends in Computer Graphics and Vision 6, 185–365.

Oliver, A., Freixenet, J., Marti, J., Perez, E., Pont, J., Denton, E.R., Zwiggelaar, R., 2010. A review of automatic mass detection and segmentation in mammographic images. Medical Image Analysis 14, 87–110.

Rahmati, P., Adler, A., Hamarneh, G., 2012. Mammography segmentation with maximum likelihood active contours. Medical image analysis 16, 1167–1186.

Rojas Domínguez, A., Nandi, A.K., 2009. Toward breast cancer diagnosis based on automated segmentation of masses in mammograms. Pattern Recognition 42, 1138–1148.

Sahiner, B., Chan, H.P., Petrick, N., Helvie, M.A., Hadjiiski, L.M., 2001. Improvement of mammographic mass characterization using spiculation measures and morphological features. Medical Physics 28, 1455–1465.

Sampat, M.P., Bovik, A.C., Whitman, G.J., Markey, M.K., 2008. A model-based framework for the detection of spiculated masses on mammography. Medical physics 35, 2110–2123.

Sethian, J.A., 1999. Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science. volume 3. Cambridge university press.

Shi, J., Sahiner, B., Chan, H.P., Ge, J., Hadjiiski, L., Helvie, M.A., Nees, A., Wu, Y.T., Wei, J., Zhou, C., et al., 2008. Characterization of mammographic masses based on level set segmentation with new image features and patient information. Medical physics 35, 280–290.

Shi, J., Sahiner, B., Chan, H.P., et al., 2007. Characterization of mammographic masses based on level set segmentation with new image features and patient information. Medical physics 35, 280–290.

Song, E., Jiang, L., Jin, R., Zhang, L., Yuan, Y., Li, Q., 2009. Breast mass segmentation in mammography using plane fitting and dynamic programming. Academic radiology 16, 826–835.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15, 1929–1958.

Tang, J., Rangayyan, R.M., Xu, J., El Naqa, I., Yang, Y., 2009. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. Information Technology in Biomedicine, IEEE Transactions on 13, 236–251.

Timp, S., Karssemeijer, N., 2004. A new 2d segmentation method based on dynamic programming applied to computer aided detection in mammography. Medical Physics 31, 958–971.

Varela, C., Timp, S., Karssemeijer, N., 2006. Use of border information in the classification of mammographic masses. Physics in Medicine and Biology 51, 425.

Wainwright, M.J., Jaakkola, T.S., Willsky, A.S., 2003. Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching, in: Workshop on Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics Np. p. 97.

Wei, J., Sahiner, B., Hadjiiski, L.M., Chan, H.P., Petrick, N., Helvie, M.A., Roubidoux, M.A., Ge, J., Zhou, C., 2005. Computer-aided detection of breast masses on full field digital mammograms. Medical physics 32, 2827–2838.

Yu, M., Huang, Q., Jin, R., Song, E., Liu, H., Hung, C.C., 2012. A novel segmentation method for convex lesions based on dynamic programming with local intra-class variance, in: Proceedings of the 27th Annual ACM Symposium on Applied Computing, ACM. pp. 39–44.

Zhang, Y., Sohn, K., Villegas, R., Pan, G., Lee, H., 2015. Improving object detection with deep convolutional networks via bayesian optimization and structured prediction, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 249 – 258.

Chapter 10

Conclusion and Future Works

Automated detection, segmentation and classification of masses in mammograms represent the essential steps in a CAD system that can act as a second reader in breast cancer screening programs. CAD systems can help radiologists increase their sensitivity and specificity in the screening of breast cancer if those systems detect, segment and classify breast lesions robustly and accurately. Building a robust and accurate CAD system for the automatic analysis of masses in mammograms is difficult because of the low signal to noise ratio of masses in comparison with surrounding tissues, lack of consistent shape and appearance patterns of masses, and limited availability of annotated public datasets. The methodologies proposed in this thesis addresses the problem of automated detection, segmentation and classification of masses using public datasets and we consider that our methodologies achieve state-of-the-art results for all these three problems. We show the result of detection, segmentation and classification of masses using five-fold cross validation experiments on INbreast dataset, which proves that our results are robust. We have also proposed a fully automated CAD system for the analysis of masses in mammograms, which can also act as a baseline result for future CAD systems.

In this chapter, we discuss the main contributions of our work, its limitations and future works.

10.1 Summary of Contributions

In this thesis, we proposed a combination of several machine learning techniques with deep learning models to build a fully automated system for the detection, segmentation and classification of masses from mammograms. We tested these techniques with manual and fully-automated settings using two publicly available datasets. We show that we can achieve a state-of-the-art results using our proposed methodologies. The main contributions of this thesis can be summarised as:

1. In Chapter 4, 5 and 6, we introduce novel methodologies for the breast mass segmentation problem using two structured output prediction models, namely SSVM and CRF. These two models combine several deep learning models. The inference with the SSVM model is based on graph cuts and the CRF is based on TRW. The parameters of SSVM model are learned with the cutting plane algorithm whereas CRF uses truncated fitting. Experiments show that our methodologies for breast mass segmentation produce state-of-the-art results in the public datasets INbreast [77] and DDSM-BCRP [78];

2. In Chapter 7, we focus on the mass detection problem and formulate it as a mass bounding box detection problem. The main novelty of our mass detection approach is that we use several stages of deep learning and random forest classifiers in a cascade model. The first stage consists of m-DBN and GMM for the detection of mass candidate regions, where the goal is to have 100% detection at the expense of a large false positive rate per image. In later stages, we use two stages of R-CNN and two stages of RF classifiers to reduce the false positive detections, producing a true positive rate of 0.9 with less than one false positive per image. Our mass detection algorithm produces state-of-the-result compared to existing methodologies for the breast mass detection problem (see Table 1 in Chapter 7);

3. In Chapter 8, we propose a transfer learning approach for the classification of masses into malignant or benign. Our transfer learning approach for breast mass classification comprises two stages. The first stage is the pre-training of the CNN model which regresses the values of hand-crafted features. The pre-trained CNN model is fine tuned using the class labels in the second stage to produce a mass classification system. We then classify the features from the fully connected layer of the CNN model using a RF classifier. Our methodology for breast mass classification produces the state-of-the-art result in the INbreast [77] dataset (see Table 1 in Chapter 8); and

4. In Chapter 9, we propose a fully automated CAD system for the analysis of masses from mammograms. Our mass detection refinement uses a local search algorithm based on Bayesian optimisation to refine the mass detected from the cascade of CNN and RF classifiers. The refined detection is used by the mass segmentation module. We then refine the mass segmentation with an active contour model. The refined segmentation and detection bounding boxes are used as input to the breast mass classification module described in Chapter 8. These steps form a complete fully automated CAD system for the analysis of masses from mammograms. We have also shown the results of the detection refinement step in Fig. 7 and fully automated segmentation and classification results in Table. 2 and 3 of Chapter 9.

10.2 Future Work

Our method successfully shows how deep learning model can be combined with other machine learning techniques for the detection, segmentation and classification of masses from mammograms. However, we believe that there are some innovations that could improve our methodology, as follows:

1. The results from the experiments concerning mass detection in Chapter 7 show that our methodology produces a higher number of false positives per image in the dataset containing digitised film samples (DDSM-BCRP) compared to FFDM samples (INbreast). This might be related to the fact that we use an identical pre-processing method for both datasets, which seems to work well with the FFDM dataset containing high signal to noise ratio (SNR), but not with digitised film dataset with low SNR. Therefore, we plan to develop a novel pre-processing methodology that can work with FFDM dataset and digitised film based datasets;

2. The other issue with our mass detection methodology, presented in Chapter 7 is that it appears that we are overfitting the training set, which is due to small training sets that we use in this thesis. This is a common issue with most of medical imaging problem, but more recently, the Breast Cancer Digital Repository (BCDR) dataset has been made public containing larger datasets of mammograms [135], which can address this overfitting issue;

3. Similarly, the visual results of segmentation in Chapters 4, 5 and 6 show that our segmentation methodologies produce smoother results compared to the ground truth annotation. This is due to the fact that the majority of mass annotations in INbreast and DDSM-BCRP datasets are round and oval shaped. This may have negative effect on segmentation and classification results of malignant masses characterised by a radiating spiculated shape. We plan to incorporate star shaped priors and train the model with the BCDR dataset [135] containing enough such cases in future;

4. Our current methodology for mass detection and classification is based on information from single view, but we can incorporate multiple view information so as to further reduce the number of false positives per image during the detection process;

5. Detection of micro-calcifications also plays an important part in the diagnosis of the patient, so we will build a system that can incorporate the analysis of both masses and calcifications;

6. We will build an end-to-end system similar to Fast R-CNN [136] for the detection, segmentation and classification tasks, which have been shown to produce more accurate results in the visual object detection problem;

7. Finally, we will apply our methodology for other problems in medical image analysis.

Bibliography

[1] Ahmedin Jemal, Rebecca Siegel, Elizabeth Ward, Yongping Hao, Jiaquan Xu, Taylor Murray, and Michael J Thun, "Cancer statistics, 2008," *CA: a cancer journal for clinicians*, vol. 58, no. 2, pp. 71–96, 2008.

[2] Lazio Tabar, A Gad, LH Holmberg, U Ljungquist, Kopparberg County Project Group, CJG Fagerberg, L Baldetorp, O Gröntoft, B Lundström, JC Månson, et al., "Reduction in mortality from breast cancer after mass screening with mammography: randomised trial from the breast cancer screening working group of the swedish national board of health and welfare," *The Lancet*, vol. 325, no. 8433, pp. 829–832, 1985.

[3] Edward A Sickles, "Breast cancer screening outcomes in women ages 40-49: clinical experience with service screening using modern mammography.," *Journal of the National Cancer Institute. Monographs*, , no. 22, pp. 99–104, 1996.

[4] Arnau Oliver, Jordi Freixenet, Joan Marti, Elsa Perez, Josep Pont, Erika RE Denton, and Reyer Zwiggelaar, "A review of automatic mass detection and segmentation in mammographic images," *Medical Image Analysis*, vol. 14, no. 2, pp. 87–110, 2010.

[5] Jinshan Tang, Rangaraj M Rangayyan, Jun Xu, Issam El Naqa, and Yongyi Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 13, no. 2, pp. 236–251, 2009.

[6] C Dromain, B Boyer, R Ferre, S Canale, S Delaloge, and C Balleyguier, "Computed-aided diagnosis (cad) in the detection of breast cancer," *European journal of radiology*, vol. 82, no. 3, pp. 417–423, 2013.

[7] I Anttinen, M Pamilo, M Soiva, and M Roiha, "Double reading of mammography screening films-one radiologist or two?," *Clinical Radiology*, vol. 48, no. 6, pp. 414–421, 1993.

[8] Maryellen L Giger and AN Pritzker, "Medical imaging and computers in the diagnosis of breast cancer," in *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, 2014, pp. 918908–918908.

[9] Robert M Nishikawa, "Current status and future directions of computer-aided diagnosis in mammography," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4, pp. 224–235, 2007.

[10] Heang-Ping Chan, Kunio Doi, CARL J VYBRONY, Robert A Schmidt, Charles E Metz, Kwok Leung Lam, Toshihiro Ogura, Yuzheng Wu, and Heber MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis.," *Investigative radiology*, vol. 25, no. 10, pp. 1102–1110, 1990.

[11] ML Giger, JM Boone, and HP Chan, "History and status of cad and quantitative image analysis," *Medical Physics*, 2008.

[12] HD Cheng, XJ Shi, Rui Min, LM Hu, XP Cai, and HN Du, "Approaches for automated detection and classification of masses in mammograms," *Pattern recognition*, vol. 39, no. 4, pp. 646–668, 2006.

[13] Victor G Martinez, Daniel M Gamo, Juan Rios, and Amparo Vilarrasa, "Iterative method for automatic detection of masses in digital mammograms for computer-aided diagnosis," in *Medical Imaging'99*. International Society for Optics and Photonics, 1999, pp. 1086–1093.

[14] D Brzakovic, XM Luo, and P Brzakovic, "An approach to automated detection of tumors in mammograms," *Medical Imaging, IEEE Transactions on*, vol. 9, no. 3, pp. 233–241, 1990.

[15] Tomoka Matsubara, Hiroshi Fujita, Satoshi Kasai, Miki Goto, Yoshinobu Tani, Takeshi Hara, and Tokiko Endo, "Development of new schemes for detection and analysis of mammographic masses," in *Intelligent Information Systems, 1997. IIS'97. Proceedings*. IEEE, 1997, pp. 63–66.

[16] Maria Kallergi, Kevin Woods, Laurence P Clarke, Wei Qian, and Robert A Clark, "Image segmentation in digital mammography: comparison of local thresholding and region growing algorithms," *Computerized medical imaging and graphics*, vol. 16, no. 5, pp. 323–331, 1992.

[17] Ehsan Kozegar, Mohsen Soryani, Behrouz Minaei, Inês Domingues, et al., "Assessment of a novel mass detection algorithm in mammograms," *Journal of cancer research and therapeutics*, vol. 9, no. 4, pp. 592, 2013.

[18] Nicholas Petrick, Heang-Ping Chan, Berkman Sahiner, and Datong Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," *Medical Imaging, IEEE Transactions on*, vol. 15, no. 1, pp. 59–67, 1996.

[19] Nicholas Petrick, Heang-Ping Chan, Berkman Sahiner, and Mark A Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," *Medical physics*, vol. 26, no. 8, pp. 1642–1654, 1999.

[20] Hidefumi Kobatake and Shigeru Hashimoto, "Convergence index filter for vector fields," *Image Processing, IEEE Transactions on*, vol. 8, no. 8, pp. 1029–1038, 1999.

[21] Hidefumi Kobatake and Yukiyasu Yoshinaga, "Detection of spicules on mammogram based on skeleton analysis," *Medical Imaging, IEEE Transactions on*, vol. 15, no. 3, pp. 235–245, 1996.

[22] Guido M Te Brake and Nico Karssemeijer, "Single and multiscale detection of masses in digital mammograms," *Medical Imaging, IEEE Transactions on*, vol. 18, no. 7, pp. 628–639, 1999.

[23] William E Polakowski, Donald A Cournoyer, Steven K Rogers, Martin P DeSimio, Dennis W Ruck, Jeffrey W Hoffmeister, and Richard A Raines, "Computer-aided breast cancer detection and diagnosis of masses using difference of gaussians and derivative-based feature saliency," *Medical Imaging, IEEE Transactions on*, vol. 16, no. 6, pp. 811–819, 1997.

[24] Lori M Bruce and Reza R Adhami, "Wavelet-based feature extraction for mammographic lesion recognition," in *Medical Imaging 1997*. International Society for Optics and Photonics, 1997, pp. 779–789.

[25] Guido M Te Brake and Nico Karssemeijer, "Segmentation of suspicious densities in digital mammograms," *Medical physics*, vol. 28, no. 2, pp. 259–266, 2001.

[26] Wiro J Niessen, Bart M Romeny, and Max A Viergever, "Geodesic deformable models for medical image analysis," *Medical Imaging, IEEE Transactions on*, vol. 17, no. 4, pp. 634–641, 1998.

[27] Michael Kass, Andrew Witkin, and Demetri Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[28] Stanley Osher and James A Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations," *Journal of computational physics*, vol. 79, no. 1, pp. 12–49, 1988.

[29] Tony F Chan, Luminita Vese, et al., "Active contours without edges," *Image processing, IEEE transactions on*, vol. 10, no. 2, pp. 266–277, 2001.

[30] Peyman Rahmati, Andy Adler, and Ghassan Hamarneh, "Mammography segmentation with maximum likelihood active contours," *Medical image analysis*, vol. 16, no. 6, pp. 1167–1186, 2012.

[31] Yong Jin Lee, Jeong Mi Park, and Hyun Wook Park, "Mammographic mass detection by adaptive thresholding and region growing," *International Journal of Imaging Systems and Technology*, vol. 11, no. 5, pp. 340–346, 2000.

[32] Guido M te Brake, Mark J Stoutjesdijk, and Nico Karssemeijer, "Discrete dynamic contour model for mass segmentation in digital mammograms," in *Medical Imaging'99*. International Society for Optics and Photonics, 1999, pp. 911–919.

[33] Matthew A Kupinski and Maryellen L Giger, "Automated seeded lesion segmentation on digital mammograms," *Medical Imaging, IEEE Transactions on*, vol. 17, no. 4, pp. 510–517, 1998.

[34] Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Mark A Helvie, Dorit D Adler, and Mitchell M Goodsitt, "Classification of masses on mammograms using rubber-band straightening transform and feature analysis," in *Medical Imaging 1996*. International Society for Optics and Photonics, 1996, pp. 44–50.

[35] Zhimin Huo, Maryellen L Giger, Carl J Vyborny, Dulcy E Wolverton, and Charles E Metz, "Computerized classification of benign and malignant masses on digitized mammograms: a study of robustness," *Academic Radiology*, vol. 7, no. 12, pp. 1077–1084, 2000.

[36] Berkman Sahiner, Nicholas Petrick, Heang-Ping Chan, Lubomir M Hadjiiski, Chintana Paramagul, Mark A Helvie, and Metin N Gurcan, "Computer-aided characterization of mammographic masses: accuracy of mass segmentation and its effects on characterization," *Medical Imaging, IEEE Transactions on*, vol. 20, no. 12, pp. 1275–1284, 2001.

[37] Huai-Dong Li, Maria Kallergi, Laurence P Clarke, Vijay K Jain, and Robert A Clark, "Markov random field for tumor detection in digital mammography," *IEEE transactions on medical imaging*, vol. 14, no. 3, pp. 565–576, 1995.

[38] Mary L Comer, Sheng Liu, and Edward J Delp, "Statistical segmentation of mammograms," in *Proceedings of the 3rd International Workshop on Digital Mammography*, 1996, pp. 475–478.

[39] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley, "Deep learning and structured prediction for the segmentation of mass in mammograms," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pp. 605–612. Springer, 2015.

[40] Michael Beller, Rainer Stotzka, Tim Oliver Müller, and Hartmut Gemmeke, "An example-based system to support the segmentation of stellate lesions," in *Bildverarbeitung für die Medizin 2005*, pp. 475–479. Springer, 2005.

[41] Jun Wei, Berkman Sahiner, Lubomir M Hadjiiski, Heang-Ping Chan, Nicholas Petrick, Mark A Helvie, Marilyn A Roubidoux, Jun Ge, and Chuan Zhou, "Computer-aided detection of breast masses on full field digital mammograms," *Medical physics*, vol. 32, no. 9, pp. 2827–2838, 2005.

[42] Guido M te Brake, Nico Karssemeijer, and Jan HCL Hendriks, "An automatic method to discriminate malignant masses from normal tissue in digital mammograms," *Physics in Medicine and Biology*, vol. 45, no. 10, pp. 2843, 2000.

[43] Nevine H Eltonsy, Georgia D Tourassi, and Adel Said Elmaghraby, "A concentric morphology model for the detection of masses in mammography," *Medical Imaging, IEEE Transactions on*, vol. 26, no. 6, pp. 880–889, 2007.

[44] C Varela, S Timp, and N Karssemeijer, "Use of border information in the classification of mammographic masses," *Physics in Medicine and Biology*, vol. 51, no. 2, pp. 425, 2006.

[45] I Domingues, E Sales, JS Cardoso, and WCA Pereira, "Inbreast-database masses characterization," *XXIII Congresso Brasileiro de Engenlaria Biomedico (CBEB)*, 2012.

[46] N. Dhungel, G. Carneiro, and A.P. Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*, Nov 2015, pp. 1–8.

[47] John E Ball and Lori Mann Bruce, "Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*. IEEE, 2007, pp. 4973–4978.

[48] Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Mark A Helvie, and Lubomir M Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Medical Physics*, vol. 28, no. 7, pp. 1455–1465, 2001.

[49] James Albert Sethian, *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*, vol. 3, Cambridge university press, 1999.

[50] Jiazheng Shi, Berkman Sahiner, Heang-Ping Chan, et al., "Characterization of mammographic masses based on level set segmentation with new image features and patient information," *Medical physics*, vol. 35, no. 1, pp. 280–290, 2007.

[51] Jaime S Cardoso, Inês Domingues, and Hélder P Oliveira, "Closed shortest path in the original coordinates with an application to breast cancer," *International Journal of Pattern Recognition and Artificial Intelligence*, 2014.

[52] Alfonso Rojas Domínguez and Asoke K Nandi, "Toward breast cancer diagnosis based on automated segmentation of masses in mammograms," *Pattern Recognition*, vol. 42, no. 6, pp. 1138–1148, 2009.

[53] Enmin Song, Luan Jiang, Renchao Jin, Lin Zhang, Yuan Yuan, and Qiang Li, "Breast mass segmentation in mammography using plane fitting and dynamic programming," *Academic radiology*, vol. 16, no. 7, pp. 826–835, 2009.

[54] Sheila Timp and Nico Karssemeijer, "A new 2d segmentation method based on dynamic programming applied to computer aided detection in mammography," *Medical Physics*, vol. 31, no. 5, pp. 958–971, 2004.

[55] Mali Yu, Qiliang Huang, Renchao Jin, Enmin Song, Hong Liu, and Chih-Cheng Hung, "A novel segmentation method for convex lesions based on dynamic programming with local intra-class variance," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. ACM, 2012, pp. 39–44.

[56] Renato Campanini, Danilo Dongiovanni, Emiro Iampieri, Nico Lanconelli, Matteo Masotti, Giuseppe Palermo, Alessandro Riccardi, and Matteo Roffilli, "A novel featureless approach to mass detection in digital mammograms based on support vector machines," *Physics in Medicine and Biology*, vol. 49, no. 6, pp. 961, 2004.

[57] Mehul P Sampat, Alan C Bovik, Gary J Whitman, and Mia K Markey, "A model-based framework for the detection of spiculated masses on mammographya)," *Medical physics*, vol. 35, no. 5, pp. 2110–2123, 2008.

[58] Roberto Bellotti, Francesco De Carlo, Sonia Tangaro, Gianfranco Gargano, Giuseppe Maggipinto, Marcello Castellano, Raffaella Massafra, Donato Cascio, Francesco Fauci, Rosario Magro, et al., "A completely automated cad system for mass detection in a large mammographic database," *Medical physics*, vol. 33, no. 8, pp. 3066–3075, 2006.

[59] Berkman Sahiner, Heang-Ping Chan, Datong Wei, Nicholas Petrick, Mark A Helvie, Dorit D Adler, and Mitchell M Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue," *Medical Physics*, vol. 23, no. 10, pp. 1671–1684, 1996.

[60] Yann LeCun and Yoshua Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, 1995.

[61] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. 2012, pp. 1097–1105, Curran Associates, Inc.

[62] Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Learning hierarchical features for scene labeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, 2013.

[63] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 580–587.

[64] Yuting Zhang, Kihyuk Sohn, Ruben Villegas, Gang Pan, and Honglak Lee, "Improving object detection with deep convolutional networks via bayesian optimization and structured prediction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, june 2015, pp. 249 – 258.

[65] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pp. 411–418. Springer, 2013.

[66] Holger R Roth, Le Lu, Ari Seff, Kevin M Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, and Ronald M Summers, "A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, pp. 520–527. Springer, 2014.

[67] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pp. 652–660. Springer, 2015.

[68] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[69] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[70] Leo Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[71] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[72] Justin Domke, "Learning graphical model parameters with approximate marginal inference," *arXiv preprint arXiv:1301.3193*, 2013.

[73] Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky, “Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching,” in *Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics Np, 2003, vol. 21, p. 97.

[74] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun, “Large margin methods for structured and interdependent output variables,” in *JMLR*, 2005, pp. 1453–1484.

[75] Martin Szummer, Pushmeet Kohli, and Derek Hoiem, “Learning crfs using graph cuts,” in *Computer Vision–ECCV 2008*, pp. 582–595. Springer, 2008.

[76] Anne Jorstad and Pascal Fua, “Refining mitochondria segmentation in electron microscopy imagery with active surfaces,” in *Computer Vision-ECCV 2014 Workshops*. Springer, 2014, pp. 367–379.

[77] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S Cardoso, “Inbreast: toward a full-field digital mammographic database,” *Academic Radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[78] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and P Kegelmeyer, “The digital database for screening mammography,” in *Proceedings of the 5th international workshop on digital mammography*, 2000, pp. 212–218.

[79] Alexander Horsch, Alexander Hapfelmeier, and Matthias Elter, “Needs assessment for next generation computer-aided mammography reference image databases and evaluation studies,” *International journal of computer assisted radiology and surgery*, vol. 6, no. 6, pp. 749–767, 2011.

[80] Yuri Boykov, Olga Veksler, and Ramin Zabih, “Fast approximate energy minimization via graph cuts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 23, no. 11, pp. 1222–1239, 2001.

[81] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley, “The automated learning of deep features for breast mass classification from mammograms,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016*, p. Accepted for publication. Springer, 2016.

[82] Maryellen L Giger, Nico Karssemeijer, and Julia A Schnabel, “Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer,” *Annual review of biomedical engineering*, vol. 15, pp. 327–357, 2013.

[83] Robert Hummel, “Image enhancement by histogram transformation,” *Computer graphics and image processing*, vol. 6, no. 2, pp. 184–195, 1977.

[84] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

[85] Karel Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics gems IV*. Academic Press Professional, Inc., 1994, pp. 474–485.

[86] Shelda Sajeev, Mariusz Bajger, and Gobert Lee, "Segmentation of breast masses in local dense background using adaptive clip limit-clahe," in *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*. IEEE, 2015, pp. 1–8.

[87] John Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 6, pp. 679–698, 1986.

[88] David Marr and Ellen Hildreth, "Theory of edge detection," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 207, no. 1167, pp. 187–217, 1980.

[89] Nico Karssemeijer and Guido M te Brake, "Detection of stellate distortions in mammograms," *Medical Imaging, IEEE Transactions on*, vol. 15, no. 5, pp. 611–619, 1996.

[90] Sebastien Morrison and Laurie M Linnett, "A model based approach to object detection in digital mammography," in *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*. IEEE, 1999, vol. 2, pp. 182–186.

[91] Shuk-Mei Lai, Xiaobo Li, and WF Biscof, "On techniques for detecting circumscribed masses in mammograms," *Medical Imaging, IEEE Transactions on*, vol. 8, no. 4, pp. 377–386, 1989.

[92] Shun Leung Ng and Walter F Bischof, "Automated detection and classification of breast tumors," *Computers and Biomedical Research*, vol. 25, no. 3, pp. 218–237, 1992.

[93] Feng Liu, Fang Zhang, Zhulin Gong, Ying Chen, and Weimin Chai, "A fully automated scheme for mass detection and segmentation in mammograms," in *Biomedical Engineering and Informatics (BMEI), 2012 5th International Conference on*. IEEE, 2012, pp. 140–144.

[94] Jiazheng Shi, Berkman Sahiner, Heang-Ping Chan, Jun Ge, Lubomir Hadjiiski, Mark A Helvie, Alexis Nees, Yi-Ta Wu, Jun Wei, Chuan Zhou, et al., "Characterization of mammographic masses based on level set segmentation with new image features and patient information," *Medical physics*, vol. 35, no. 1, pp. 280–290, 2008.

[95] Fang-Fang Yin, Maryellen L Giger, Kunio Doi, Charles E Metz, Carl J Vyborny, and Robert A Schmidt, "Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images," *Medical Physics*, vol. 18, no. 5, pp. 955–963, 1991.

[96] Tin-Kit Lau and Walter F Bischof, "Automated detection of breast tumors using the asymmetry approach," *Computers and biomedical research*, vol. 24, no. 3, pp. 273–295, 1991.

[97] Saskia van Engeland and Nico Karssemeijer, "Exploitation of correspondence between cc and mlo views in computer aided mass detection," in *International Workshop on Digital Mammography*. Springer, 2006, pp. 237–242.

[98] Jun Wei, Heang-Ping Chan, Berkman Sahiner, Chuan Zhou, Lubomir M Hadjiiski, Marilyn A Roubidoux, and Mark A Helvie, "Computer-aided detection of breast masses on mammograms: Dual system approach with two-view analysis," *Medical physics*, vol. 36, no. 10, pp. 4451–4460, 2009.

[99] Marina Velikova, Maurice Samulski, Peter JF Lucas, and Nico Karssemeijer, "Improved mammographic cad performance using multi-view information: a bayesian network framework," *Physics in Medicine and Biology*, vol. 54, no. 5, pp. 1131, 2009.

[100] Guy Amit, Sharbell Hashoul, Pavel Kisilev, Boaz Ophir, Eugene Walach, and Aviad Zlotnick, "Automatic dual-view mass detection in full-field digital mammograms," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 44–52.

[101] Nenad Vujovic and Dragana Brzakovic, "Establishing the correspondence between control points in pairs of mammographic images," *Image Processing, IEEE Transactions on*, vol. 6, no. 10, pp. 1388–1399, 1997.

[102] Robert Marti, Reyer Zwiggelaar, and Caroline ME Rubin, "Automatic point correspondence and registration based on linear structures," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 16, no. 03, pp. 331–340, 2002.

[103] Lionel CC Wai and Michael Brady, "Curvilinear structure based mammographic registration," in *Computer Vision for Biomedical Image Applications*, pp. 261–270. Springer, 2005.

[104] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5325–5334.

[105] Joshua J Fenton, Stephen H Taplin, Patricia A Carney, et al., "Influence of computer-aided detection on performance of screening mammography," *New England Journal of Medicine*, vol. 356, no. 14, pp. 1399–1409, 2007.

[106] N. Dhungel, G. Carneiro, and A. P. Bradley, "Deep structured learning for mass segmentation from mammograms," in *Image Processing (ICIP), 2015 IEEE International Conference on*, Sept 2015, pp. 2950–2954.

[107] N. Dhungel, G. Carneiro, and A. P. Bradley, "Tree re-weighted belief propagation using deep learning potentials for mass segmentation from mammograms," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, April 2015, pp. 760–763.

[108] David M Catarious Jr, Alan H Baydush, and Carey E Floyd Jr, "Incorporation of an iterative, linear segmentation routine into a mammographic mass cad system," *Medical physics*, vol. 31, no. 6, pp. 1512–1520, 2004.

[109] Shengzhou Xu, Hong Liu, and Enmin Song, "Marker-controlled watershed for lesion segmentation in mammograms," *Journal of digital imaging*, vol. 24, no. 5, pp. 754–763, 2011.

[110] Michael Beller, Rainer Stotzka, Tim Oliver Müller, and Hartmut Gemmeke, "An example-based system to support the segmentation of stellate lesions," in *Bildverarbeitung für die Medizin 2005*, pp. 475–479. Springer, 2005.

[111] Zhimin Huo, Maryellen L Giger, Carl J Vyborny, Ulrich Bick, Ping Lu, Dulcy E Wolverton, and Robert A Schmidt, "Analysis of spiculation in the computerized classification of mammographic masses," *Medical Physics*, vol. 22, no. 10, pp. 1569–1579, 1995.

[112] Kostas Haris, Serafim N Efstratiadis, Nikolaos Maglaveras, and Aggelos K Katsaggelos, "Hybrid image segmentation using watersheds and fast region merging," *IEEE Transactions on image processing*, vol. 7, no. 12, pp. 1684–1699, 1998.

[113] Lisa Kinnard, S-CB Lo, Erini Makariou, Teresa Osicka, Paul Wang, Matthew T Freedman, and Mohamed Chouikha, "Likelihood function analysis for segmentation of mammographic masses for various margin groups," in *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*. IEEE, 2004, pp. 113–116.

[114] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[115] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[116] Holger R Roth, Amal Farag, Le Lu, Evrim B Turkbey, and Ronald M Summers, "Deep convolutional networks for pancreas segmentation in ct imaging," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2015, pp. 94131G–94131G.

[117] Tuan Anh Ngo and Gustavo Carneiro, "Fully automated non-rigid segmentation with distance regularized level set evolution initialized and constrained by deep-structured inference," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 3118–3125.

[118] Breast Imaging Reporting, "Data system atlas," *American college of radiology*, 2003.

[119] Xiaolan Zeng, Sarah Medeiros, Susan A Wood, Sandra Stapleton, Jimmy Roehrig, Kathryn O'Shaughnessy, and Ronald A Castellino, "Computer-aided detection for mammography: improved algorithm performance with operator determined points characterized by new metrics," in *International Congress Series*. Elsevier, 2004, vol. 1268, pp. 855–860.

[120] Jasjit S Suri, Ramachandran Chandrasekhar, Nico Lanconelli, Renato Campanini, Matteo Roffilli, Ruey-Feng Chang, Yujun Guo, Radhika Sivaramakrishna, Tibor Tot, Begona Acha, et al., "The current status and likely future of breast imaging cad," *Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer*, 2006.

[121] Henry A Swett and Perry L Miller, "Icon: a computer-based approach to differential diagnosis in radiology.," *Radiology*, vol. 163, no. 2, pp. 555–558, 1987.

[122] Henry A Swett, Paul R Fisher, Aaron I Cohn, Perry L Miller, and Pradeep G Mutalik, "Expert system-controlled image display.," *Radiology*, vol. 172, no. 2, pp. 487–493, 1989.

[123] Maryellen L Giger, Kunio Doi, H MacMahon, RM Nishikawa, KR Hoffmann, CJ Vyborny, RA Schmidt, H Jia, K Abe, and X Chen, "An" intelligent" workstation for computer-aided diagnosis.," *Radiographics*, vol. 13, no. 3, pp. 647–656, 1993.

[124] Yulei Jiang, Robert M Nishikawa, Robert A Schmidt, Charles E Metz, Maryellen L Giger, and Kunio Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Academic radiology*, vol. 6, no. 1, pp. 22–33, 1999.

[125] Karla Horsch, Maryellen L Giger, Carl J Vyborny, and Luz A Venta, "Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography," *Academic radiology*, vol. 11, no. 3, pp. 272–280, 2004.

[126] Karla Horsch, Maryellen L Giger, Carl J Vyborny, Li Lan, Ellen B Mendelson, and R Edward Hendrick, "Classification of breast lesions with multimodality computer-aided diagnosis: Observer study results on an independent clinical data set 1," *Radiology*, vol. 240, no. 2, pp. 357–368, 2006.

[127] Heang-Ping Chan, Berkman Sahiner, Mark A Helvie, Nicholas Petrick, Marilyn A Roubidoux, Todd E Wilson, Dorit D Adler, Chintana Paramagul, Joel S Newman, and Sethumadavan Sanjay-Gopal, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: An roc study 1," *Radiology*, vol. 212, no. 3, pp. 817–827, 1999.

[128] Carl Edward Rasmussen, "Gaussian processes in machine learning," in *Advanced lectures on machine learning*, pp. 63–71. Springer, 2004.

[129] Nobuyuki Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[130] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[131] Miguel A Carreira-Perpinan and Geoffrey Hinton, "On contrastive divergence learning.," in *AISTATS*. Citeseer, 2005, vol. 10, pp. 33–40.

[132] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[133] Sebastian Nowozin and Christoph Lampert, "Structured learning and prediction in computer vision," *Foundations and Trends in Computer Graphics and Vision*, vol. 6, no. 3–4, pp. 185–365, 2011.

[134] Talya Meltzer, Amir Globerson, and Yair Weiss, "Convergent message passing algorithms: a unifying view," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 393–401.

[135] Miguel A Guevara López, NG Posada, Daniel C Moura, RR Polln, José M Franco Valiente, and César Suárez Ortega, "Bcdr: a breast cancer digital repository," in *15th International Conference on Experimental Mechanics*, 2012.

[136] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.