A Deep Learning Approach for the Analysis of Masses in Mammograms with Minimal User Intervention

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

[a] *Electrical and Computer Engineering, The University of British Columbia, Canada*
[b] *Australian Centre for Visual Technologies, The University of Adelaide, Australia*
[c] *ITEE, The University of Queensland, Australia*

Abstract

We present an integrated methodology for detecting, segmenting and classifying breast masses from mammograms with minimal user intervention. 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 system detects 90% of masses at 1 false positive per image, has a segmentation accuracy of around 0.85 (Dice index) on the correctly detected masses, and overall classifies masses as malignant or benign with sensitivity

*Corresponding author. This research was conducted while N. Dhungel was with the University of Adelaide.

Email address: neerajd@ece.ubc.ca (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).

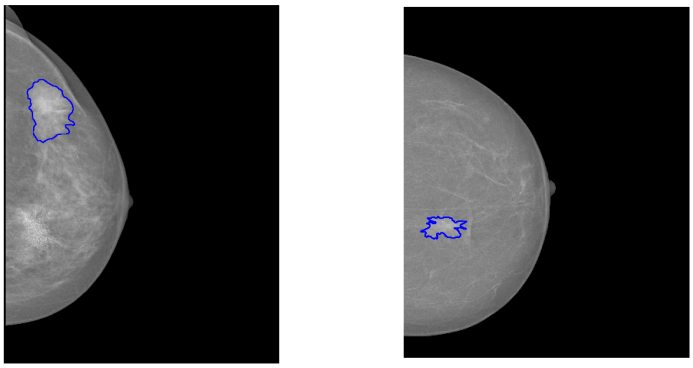
(Se) of 0.98 and specificity (Sp) of 0.7.

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 (of Health et al. (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); 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.

CAD systems are useful in the detection, segmentation and classification of breast masses, which represent challenging tasks given 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 main breast imaging modality (see Fig. 1), hinders the development and evaluation of CAD



(a) Malignant mass (b) Benign mass

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

systems. Current methodologies for mass detection (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)). generally produce 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. Moreover, recently proposed segmentation methods (Rahmati et al. (2012); Cardoso et al. (2015)) tend not to be robust to the shape and appearance variations of masses and usually have high run-time and/or memory complexities. Finally, mass classification typically uses hand-crafted features that are not optimally designed for this task (Varela et al. (2006); Shi et al. (2008); Domingues et al. (2012)).

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 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 methodology 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

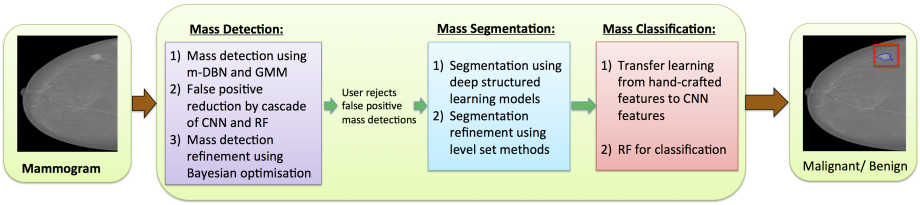


Figure 2: Our proposed methodology of breast mass detection, segmentation and classification with minimal user intervention. Mass detection is done using mass candidate generation and false positive reduction (Dhungel et al. (2015a)) with a new detection refinement. Segmentation is carried out using 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 (Dhungel et al. (2016)). The user intervention happens between the mass detection and segmentation stages, as shown in the diagram.

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 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. (2008); 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 interest-

ing 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 the underlying framework for analysing (i.e., detecting, segmenting and classifying) 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)).

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

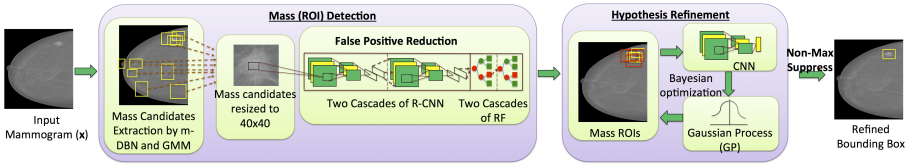


Figure 3: The proposed mass detection consists of two stages of mass ROI detection followed by hypothesis refinement. The Mass ROI detection is based on the results of m-DBN and GMM to generate candidates, followed by a false positive reduction using cascades of CNN and RF; and the hypothesis refinement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

of the mass that can be either benign(i.e., BI-RADS $\in \{2, 3\}$) or malignant (i.e., BI-RADS $\in \{4, 5, 6\}$). Note that a mammogram i without any mass annotated (i.e., no findings - BI-RADS=1) is represented by $\mathcal{A}_i = \emptyset$.

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 $f_{\mathrm{RGH}}(.)$ 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 $\{\mathbf{d}_n^*, \widetilde{\mathbf{y}}_n^*\}_{n=1}^{N_{\mathrm{RGH}}}$ 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:

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

where $f_{\mathrm{RCNN}}(.)$ is a model defined by parameters θ_{RCNN} (the weights and biases of the CNN and the linear SVM parameters), and $N_{\mathrm{RCNN}} \leq N_{\mathrm{RGH}}$ (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 $\{\mathbf{d}_n^*, \widetilde{\mathbf{y}}_n^*\}_{n=1}^{J_{\mathrm{RCNN}}}$ 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

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

where $f_{\mathrm{RF}}(.)$ represents a random forest classifier defined by parameters θ_{RF} (number of trees, number of leaves in each tree, etc.), and $N \leq N_{\mathrm{RCNN}}$ (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:

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

where $\hat{f}_N = \max_{n\in\{1,\dots,N\}} f_n$, θ_{EI} represents the parameters of model $a_{\text{EI}}(.)$, and $P(f_{N+t}|\mathbf{d}^*_{N+t}, \mathcal{B}_{N+t-1}, \theta_{\text{EI}})$ 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 intersection 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-

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 θ_{SC} for the scoring function in (4), acquisition function parameters θ_{EI} in (5), and maximum number of iterations $t_{\max}$, a threshold f_{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
$\mathcal{B}_{\text{prune}} = \{(\mathbf{d}, f) \in \mathcal{B}_{\text{new}} : f \geq f_{\text{prune}}\}$
$\mathcal{B}_{\text{ref}} = \text{NMS}(\mathcal{B}_{\text{prune}})$

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

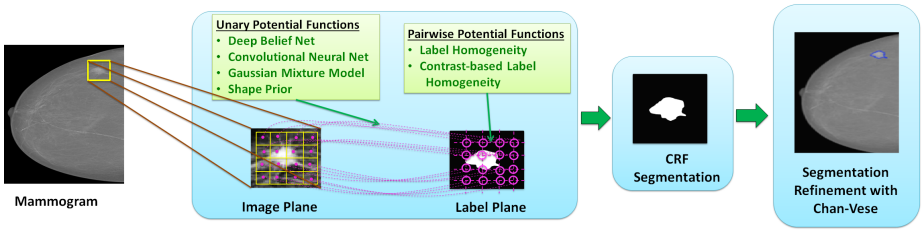


Figure 4: The proposed mass segmentation is carried out with the segmentation produced by a CRF on a low resolution image patch that is then scaled to the original image size and refined with the Chan-Vese active contour method (Chan et al. (2001)).

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} (we assume 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 Fig. 4.

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

Algorithm 2 Mass Segmentation with Refinement

Require: Mammogram $\mathbf{x}$, refined bounding box $\mathbf{d}_n \in \mathcal{B}_N$, sub-image size M_{sub}, number of iterations $t_{\max}$ for the Chan-Vese optimisation, the unary and pairwise model parameters θ_{CNNSEG}, θ_{DBNSEG}, θ_{GMMSEG}, θ_{PRIORSEG}, and structured output model θ_{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 $\mathbf{y}_n^* = \phi_{t_{\max}} \geq 0$

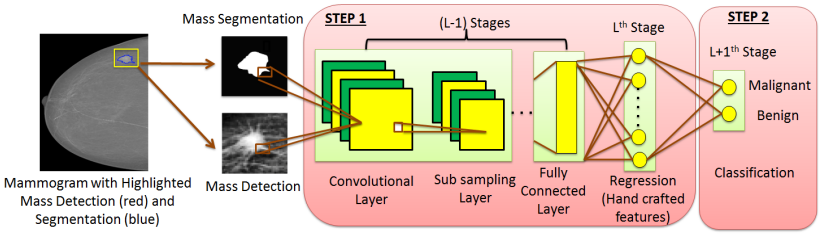


Figure 5: The proposed classification methodology consists of two steps: 1) pre-training of the CNN for regressing the values of hand-crafted features, and 2) fine-tuning the pre-trained CNN model for the mass classification problem.

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 fea-

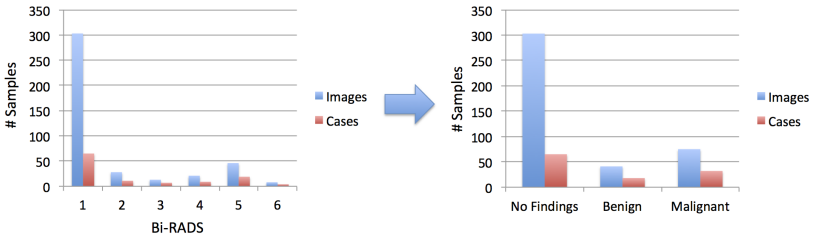


Figure 6: Distribution of images and cases on the INbreast dataset (Moreira et al. (2012)) with respect to the BI-RADS classification and the "No Findings" (BI-RADS=1), "Benign" (BI-RADS $\in \{2, 3\}$), and "Malignant" (BI-RADS $\in \{4, 5, 6\}$) classes, as defined above in Sec. 3.

tures, consisting of morphological and texture features (Varela et al. (2006)). The morphological features are computed using the segmentation mask $\mathbf{y}$, whereas 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 50 cases and 116 images have benign or malignant masses and the remaining ones do not contain any masses (i.e., "No Findings"). In particular, Fig. 6 shows how these cases and images are divided into BI-RADS and respective "No Findings" (BI-RADS=1), "Benign" (BI-RADS $\in \{2, 3\}$), and "Malignant" (BI-RADS $\in \{4, 5, 6\}$) classes, as defined above in Sec. 3. The performance of the detection methodology is assessed with all 410 images (from 115 cases), while the segmentation and classification methodologies are evaluated with 41 benign masses (from 18 cases) and 75 malignant masses (from 32 cases). In all these experiments, the cases 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 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 Set-up

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$\geq$ 0.3. This new CNN classifier defines a positive sample by IoU$\geq$ 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 Set-up

We explore a manual and a minimal user intervention set-ups for the segmentation problem, where the manual set-up relies on the manual annotations for the ROI, while the minimal user intervention set-up uses an automated ROI detection, where false positives are manually rejected (for our methodology, the automated ROI detection is produced by Algorithm 1).

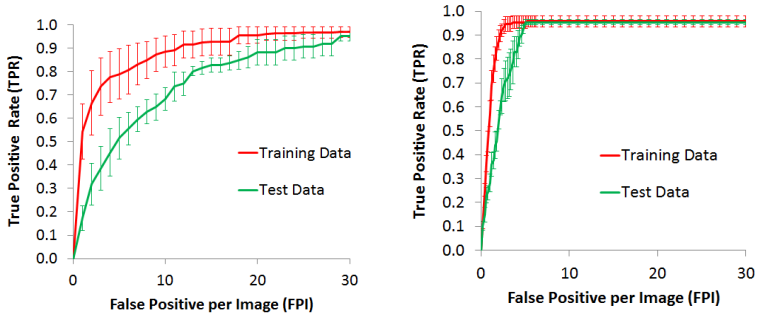
The model selection for the DBN (θ_{DBNSEG}) and CNN (θ_{CNNSEG}) unary potential functions in Algorithm. 2 is performed via cross validation using the training and validation sets. The DBN model has two layers with 200 and 500 nodes, which are trained with image patch sizes of 3×3, 5×5, and 7×7. The 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, and the following values are estimated in all folds: $\mu = 0.2, \lambda = 1, \gamma = 1$ and number of iterations $t = 10$.

6.3. Classification Experimental Set-up

We explore a manual, semi-automated and minimal user intervention set-ups for classification where the manual set-up uses the manual annotations for the

ROI *and* segmentation mask. The semi-automated set-up relies on the manual annotations for the ROI, but uses an algorithm to segment masses automatically. The minimal user intervention set-up is based on automated ROI detection and mass segmentation, where the false positive detections are manually rejected before being processed by the segmentation and classification algorithms. For our methodology, the automated ROI detection is obtained from Algorithm 1, and the segmentation is estimated from Algorithm 2.

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 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



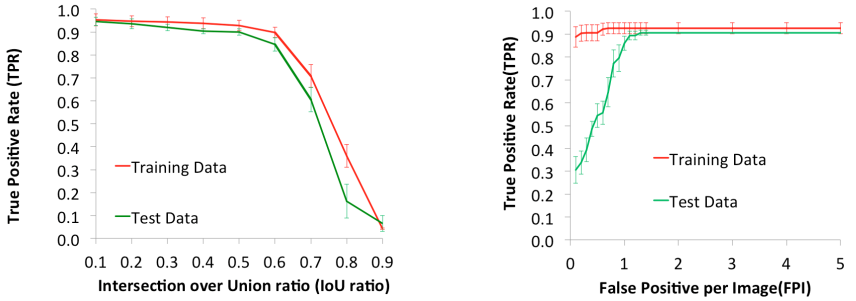
(a) FROC - Cascade of R-CNN (b) FROC - Cascade of RF

Figure 7: 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)).

on average, the RFs have 8 trees (search range in [1,1000]), each with 6 leaves (search range in [1,1000]).

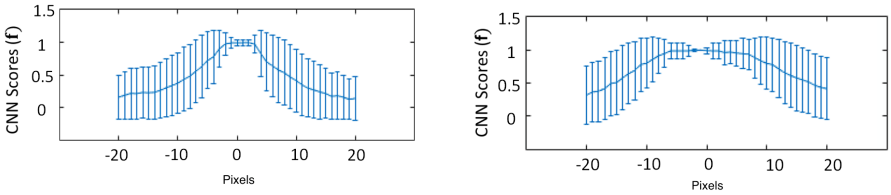
7. Experimental Results

Fig. 7-(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. 7-(b)). Figure 8-(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 with the hypothesis refinement described in Sec. 3.3. From the FROC curve in Fig. 8-(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.



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

Figure 8: 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 9: Plot of the CNN classifier in (5) as a function of the annotated bounding box horizontal (a) and vertical (b) translation.

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. 9-(a) and Fig. 9-(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. 9 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 in terms of the Dice index for training and testing data from the detected and refined ROIs from Algorithm. 1 (after false positives have been manually re-

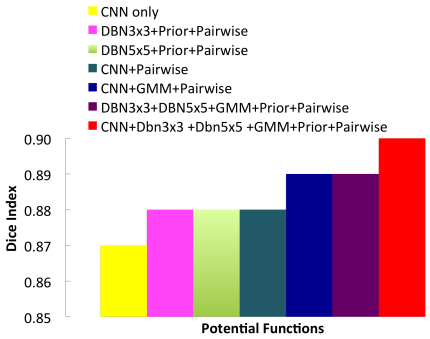


Figure 10: Effect of adding different potential functions into our CRF model (Dhungel et al. (2015b)) on the testing set of INbreast taking a manually detected ROI breast mass.

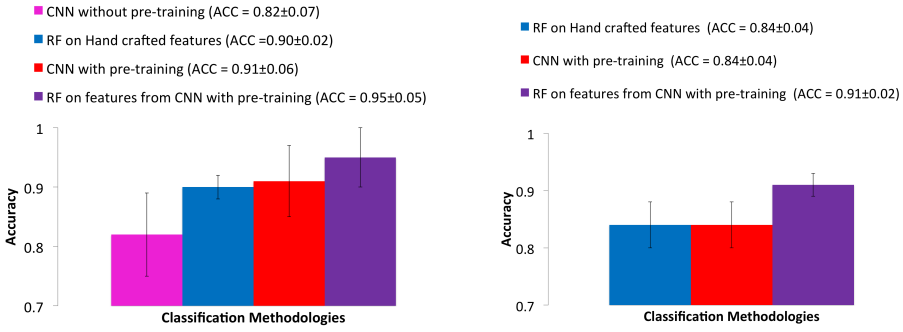
Table 1: Results of our minimal user intervention segmentation algorithm on the INbreast dataset.

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

jected). The segmentation was carried out using the combination of several potential functions (CNN+DBN3$\times$3 + DBN5$\times$5 + GMM + Prior + 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. 10 (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 initiali-

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

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

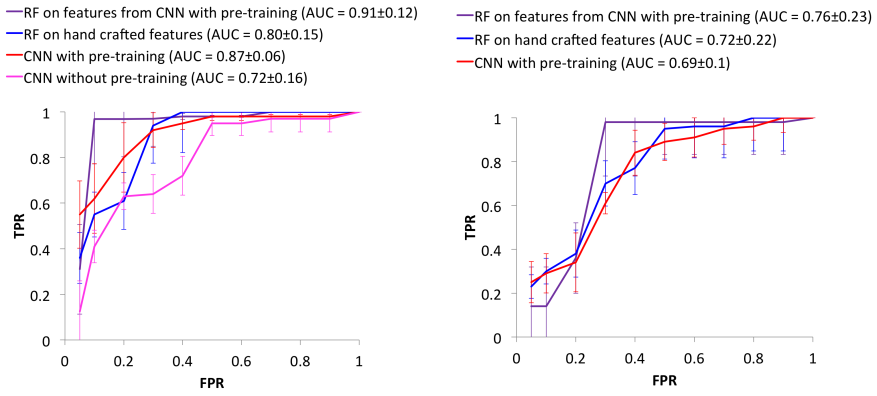


(a) Manual set-up (b) Min. user interact. set-up

Figure 11: Accuracy of various classifiers on features extracted using the methodologies described in this paper based on the manual and minimal user intervention on test data.

sation 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 "Set-up" indicates whether the mass ROI detection is performed with minimal user intervention (i.e., an automated mass detection, followed by a manual rejection of false positives), or manually (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 11-(a-b) displays the classification accuracy for both manual and automated set-



(a) Manual set-up (b) Min. user interact. set-up

Figure 12: ROC curve of various classifiers on features extracted using the methodologies described in this paper based on the manual and minimal user intervention on test data.

ups, 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 minimal user intervention set-up. In addition, we compare the results between the various models in terms of area under the ROC curve (AUC) in Figures 12-(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 minimal user intervention set-up. We also compare our classification method with other state-of-the-art methods in Tab. 3 in terms of classification accuracy (ACC) and AUC where applicable.

The total running time for our minimal user intervention system is 41 seconds per 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. 13 for the minimal user intervention detection and segmentation results and in Fig. 14 for the minimal user intervention detection, segmentation and classification system.

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

Methodology	Dataset	Set-up	ACC	AUC
Proposed RF on CNN with pre-training	INbreast	Min. user interact.	0.91 ± 0.02	0.76 ± 0.23
Proposed CNN with pre-training	INbreast	Min. user interact.	0.84 ± 0.04	0.69 ± 0.10
Proposed RF on hand-crafted features	INbreast	Min. user interact.	0.84 ± 0.04	0.72 ± 0.22
Proposed RF on CNN with pre-training	INbreast	Manual	0.95 ± 0.05	0.91 ± 0.12
Proposed CNN with pre-training	INbreast	Manual	0.91 ± 0.06	0.87 ± 0.06
Proposed RF on hand-crafted features	INbreast	Manual	0.90 ± 0.02	0.80 ± 0.15
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. (2008)	Uni. of Michigan	Semi-automated	0.83 ± 0.02	0.85 ± 0.02

8. Discussion

The results from the Fig. 8-(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. 7 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 (at a TPR ≥ 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. 10 (Dhungel et al. (2015b)) on manual set-up shows that the combination of all the potential functions (CNN + DBN3x3 + DBN5x5 + GMM + prior + pairwise) is crucial for producing state-of-the-art

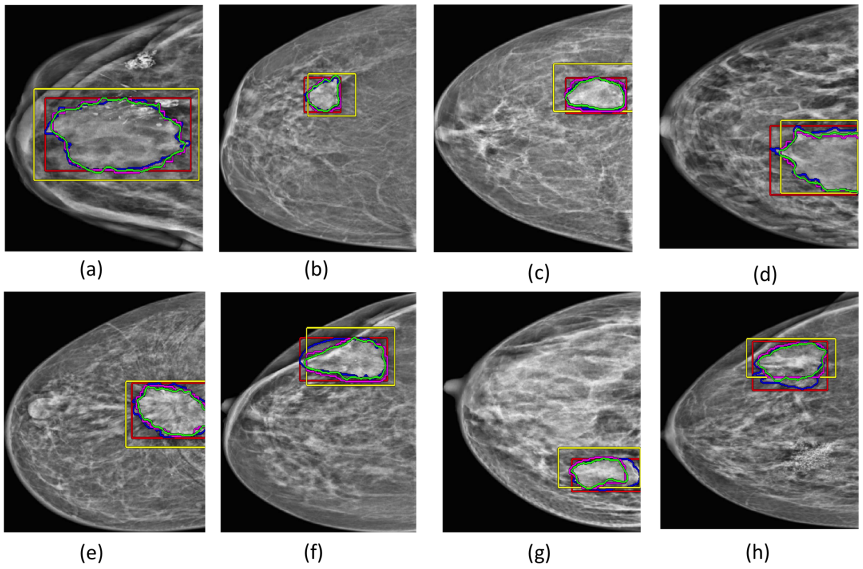


Figure 13: Examples of the minimal user intervention mass detection and segmentation with refinement. The contour with the blue line represents the ground truth annotation, red line denotes the manual ROI, yellow is the detected and refined ROI from our methodology, magenta is the segmentation from the CRF model with nearest neighbor interpolation, and green is the segmentation refined by the active contour model.

results. Therefore, we use all these potential functions in our CRF segmentation model for the minimal user intervention set-up. 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 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).

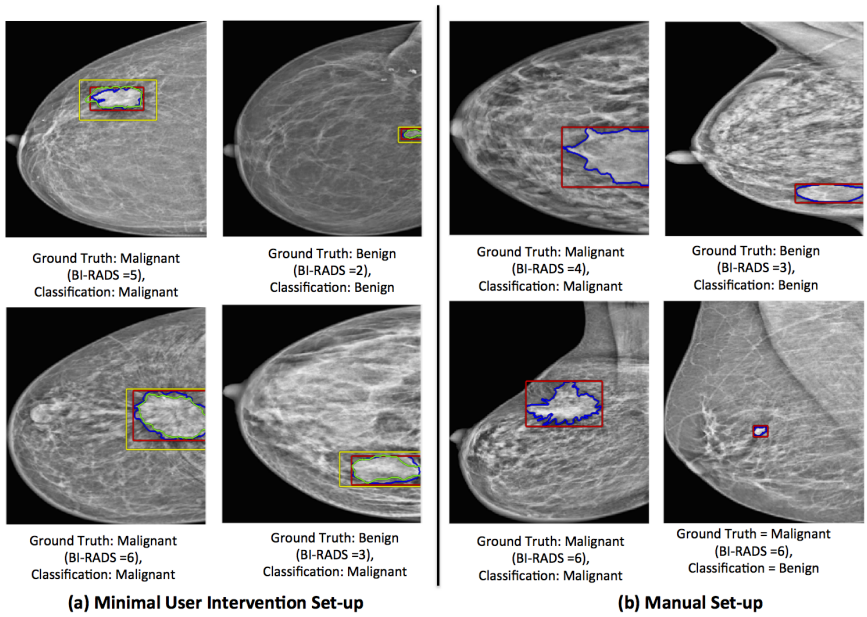


Figure 14: Examples of mass classification based on the RF model on features from CNN with pre-training using the minimal user intervention set-up and manual set-up. Red contours denote manual detection and blue denotes the manual segmentation whereas yellow contours represent the automated detection and green is the automated segmentation. Ground truth and automated classification results are shown below each image.

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 ± 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))). Moreover, it is important to explain that the better performance of the manual set-up, compared to the minimal user intervention set-up in Table 2, is due to the better alignment of the masses in the ROI provided by the manual set-up.

For the mass classification problem, the results in Figures 11 and 12 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 11 and 12 also show that the RF classifier performs better than the CNN classifier in both minimal user intervention and

manual set-ups. Here, we did not show the classification results of CNN without pre-training for the minimal user intervention system because of its poor performance on manual set-up. 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 $= 0.94 \pm 0.06$) and testing accuracy (ACC $= 0.95 \pm 0.05$) on manual set-up correlates well with each other implying good generalisation of RF on CNN features with pre-training. From the Fig.12 (a-b), we see that there is an increase in FPR and decrease in the AUC value in the minimal user intervention system compared to the manual set-up which is expected due to the better alignment of the masses in the ROI in the manual set-up. Table. 3 shows that our methodology produces competitive results, with respect to other works in the literature, in terms of classification accuracy in manual, semi-automated and minimal user intervention set-ups. The visual results in Fig. 14-(a) shows classification results using minimal user intervention set-up and Fig. 14-(b) shows the results from the manual set-up. The visual results for the minimal user intervention 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 sensitivity of 0.97 and specificity of 0.90, while the minimal user intervention set-up produces a sensitivity of 0.98 and specificity of 0.70, which shows that our proposed CAD system is robust to false positives and false negatives.

9. Future Work

In the future, we would like to build a end-to-end system capable of the detection, segmentation and classification of breast masses using a single integrated module similar to that of Fast R-CNN (Girshick (2015)) that has produced state-of-the-art result recently in the field of object detection. We would also like to try better segmentation models, such as the fully convolutional neural

networks (FCN) (Long et al. (2015)) and the U-net (Ronneberger et al. (2015)), which have produced state-of-the-art segmentation results in several computer vision datasets. In addition, we plan to apply this methodology to other similar problems involving different imaging modalities, such as mass analysis from breast magnetic resonance imaging (Gilhuijs et al. (1998)), nodule analysis from chest x-ray (Ngo and Carneiro (2015); Van Ginneken et al. (2001)), and micro-calcification analysis from mammograms (Lu et al. (2016); Yu and Guan (2000)). Finally, perhaps the most important criticism about our work is the fact that we use such small dataset to train and test the proposed methodologies. We believe that once the field acquires and makes publicly available large mammogram datasets, data will "speak for itself", and we will no longer require priors (such as CRF for segmentation from Sec. 4) or training regularisation methods (such as the use of hand-crafted features to pre-train the classifier from Sec. 5), and efforts will be shifted from the generalisation of models to the efficient processing of very large datasets. Therefore, we plan to work towards the acquisition and annotation of a large annotated dataset of mammograms, and we encourage the field to works towards this direction.

10. Conclusions

In this paper, we describe a complete minimal user intervention 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 classic 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 mass detection, segmentation and classification systems produce promising results, which can be used as baseline. We also believe that our current methodology can be incorporated in the clinical set-up as a second reader for radiologists.

References

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.

Gilhuijs, K.G., Giger, M.L., Bick, U., 1998. Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging. Medical physics 25, 1647–1654.

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.

of Health, A.I., Welfare, Australia, C., 2012. Breast cancer in australia: an overview. Cancer series no. 71. Cat. no. CAN 67, Canberra: AIHW .

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.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Lu, Z., Carneiro, G., Dhungel, N., Bradley, A.P., 2016. Automated detection of individual micro-calcifications from mammograms using a multi-stage cascade approach. arXiv preprint arXiv:1610.02251 .

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.

Ngo, T.A., Carneiro, G., 2015. Lung segmentation in chest radiographs using distance regularized level set and deep-structured learning and inference, in: Image Processing (ICIP), 2015 IEEE International Conference on, IEEE. pp. 2140–2143.

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.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 234–241.

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.

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.

Van Ginneken, B., Romeny, B.T.H., Viergever, M.A., 2001. Computer-aided diagnosis in chest radiography: a survey. IEEE Transactions on medical imaging 20, 1228–1241.

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.

Yu, S., Guan, L., 2000. A cad system for the automatic detection of clustered microcalcifications in digitized mammogram films. IEEE transactions on medical imaging 19, 115–126.

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.