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).

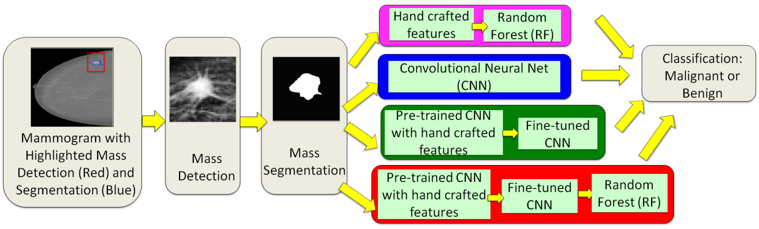

Mammogram with Highlighted Mass Detection (Red) and Segmentation (Blue)
Mass Detection
Mass Segmentation
Hand crafted features
Random Forest (RF)
Convolutional Neural Net (CNN)
Pre-trained CNN with hand crafted features
Fine-tuned CNN
Pre-trained CNN with hand crafted features
Fine-tuned CNN
Random Forest (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 $\mathbf{x} : \Omega \to \mathbb{R}$ 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, $\mathbf{s}(i)_j : \Omega \to \{0, 1\}$ 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

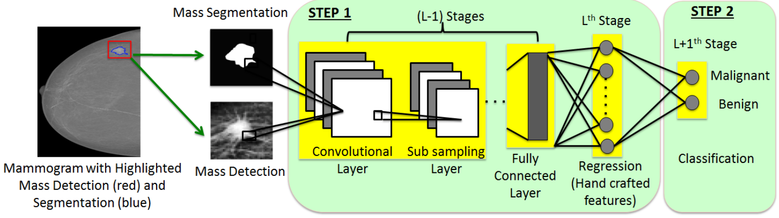


Fig. 2: Two steps of the proposed model with the pre-training of the CNN with the regression to the hand-crafted features (step 1), followed by the fine-tuning using the mass classification problem (step 2).

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:

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

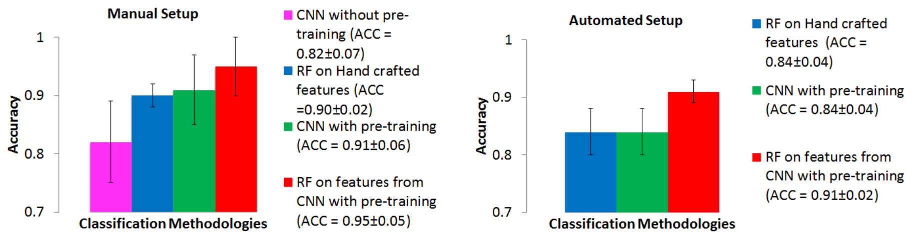
where i indexes the training images, j indexes the masses in each training image, and $\mathbf{z}^{(H)}_{(i,j)}$ denotes the vector of hand-crafted features from mass j and image i. This first step acts as a regulariser for the classifier that is sub-sequentially fine-tuned (see step 2 in Fig. 2).

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

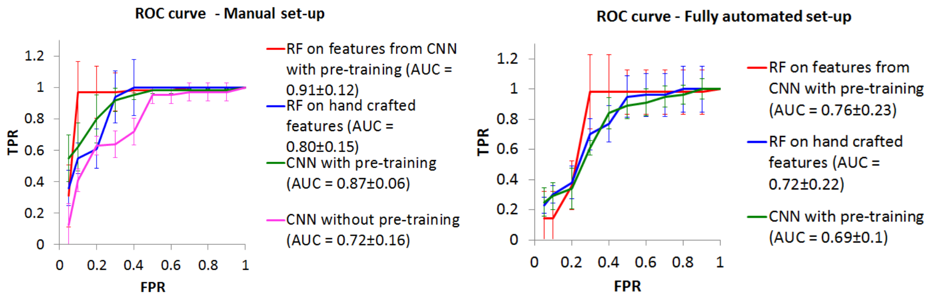
We use the publicly available INbreast dataset [10] that contains 115 cases with 410 images, where 116 images contain benign or malignant masses. Experiments are run using five fold cross validation by randomly dividing the 116 cases in a mutually exclusive manner, with 60% of the cases for training, 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 $= 0.93 \pm 0.05$ 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

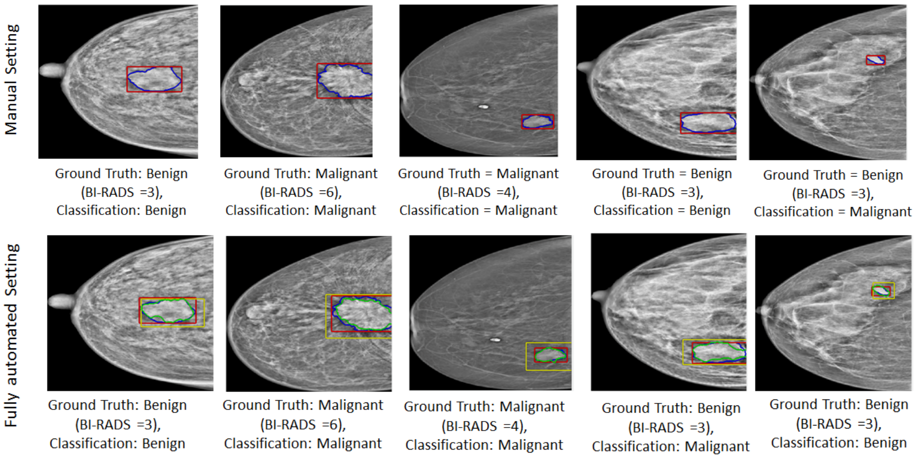

Manual Setting
Ground Truth: Benign (BI-RADS =3), Classification: Benign
Ground Truth: Malignant (BI-RADS =6), Classification: Malignant
Ground Truth = Malignant (BI-RADS =4), Classification = Malignant
Ground Truth = Benign (BI-RADS =3), Classification = Benign
Ground Truth = Benign (BI-RADS =3), Classification = Malignant
Fully automated Setting
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)