CHAPTER 1

Deep Learning Models for Classifying Mammogram Exams Containing Unregistered Multi-view Images and Segmentation Maps of Lesions[1]

Gustavo Carneiro[,a,*], Jacinto Nascimento[**] and Andrew P. Bradley[†]
* University of Adelaide, Australian Centre for Visual Technologies, North Terrace, Ingkarni Wardli Building, Adelaide, SA 5005, Australia
** Instituto Superior Técnico, Institute for Systems and Robotics, Av. Rovisco Pais, Torre Norte, 1049-001 Lisbon, Portugal
† University of Queensland, School of Information Technology and Electrical Engineering, Brisbane QLD 4072, Australia
[a] Corresponding: gustavo.carneiro@adelaide.edu.au

Abstract

In this chapter, we show two discoveries learned from the application of deep learning methods to the problem of classifying mammogram exams containing multi-view images and segmentation maps of breast lesions (i.e., masses and micro-calcifications). We first demonstrate the efficacy of pre-training a deep learning model using extremely large computer vision training sets, and then fine-tuning this same model for the classification of mammogram exams. We also show that the multi-view mammograms and segmentation maps do not need to be registered in order to produce accurate classification results using the fine-tuned deep learning model above. In particular, we take a deep learning model pre-trained to identify Imagenet classes from real images, and fine-tune it with cranio-caudal (CC) and medio-lateral oblique (MLO) mammography views of a single breast and their respective mass and micro-calcification segmentation maps in order to estimate the patients risk of developing breast cancer. This methodology is tested on two publicly available datasets (In-Breast and DDSM), and we show that our approach produces a volume under ROC surface of over 0.9 and an area under ROC curve (for a 2-class problem: benign and malignant) of over 0.9. These results show that our method can produce state-of-the-art classification results using a new comprehensive way of tackling medical image analysis problems.

Deep learning, Mammogram, Multi-view classification, Transfer learning

[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). This work is an extension of the paper published by the same authors at the Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015) [1].

© Elsevier Ltd.
All rights reserved.

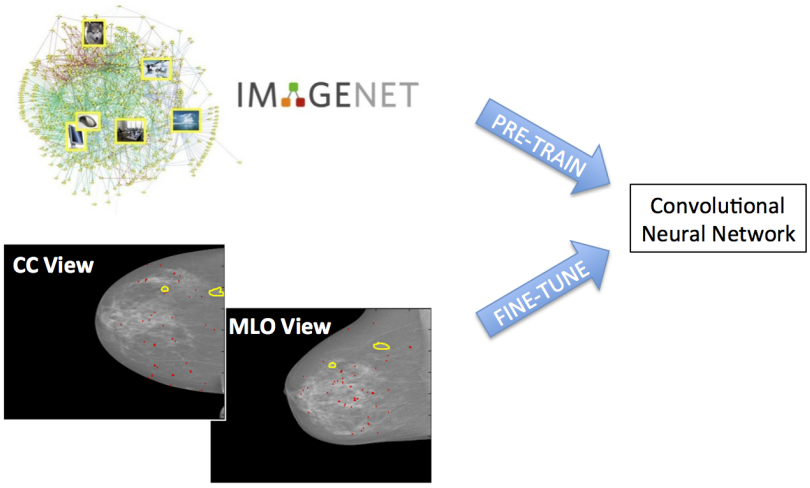

Figure 1.1 The first goal of the paper is to show how to fine-tune deep learning models, pre-trained with the computer vision database Imagenet [2], for the joint analysis of the cranio-caudal (CC) and medio-lateral oblique (MLO) mammography views. Note that the yellow annotations denote breast masses and red annotations represent micro-calcifications.

Chapter points

- Shows how to fine-tune deep learning models, pre-trained with computer vision databases, for the analysis of mammograms.
- Demonstrates how high-level deep learning model features can be used in multi-view mammogram classification problems without pre-registering the mammograms.

1. Introduction

According to recent statistical data published by the World Health Organisation (WHO), breast cancer accounts for 23% of all cancer related cases and 14% of all cancer related deaths amongst women worldwide [3]. The early detection of breast cancer in asymptomatic women using breast screening mammography is currently the most effective tool to reduce the morbidity and mortality associated with breast cancer [4]. A breast screening exam typically consists of two mammography views of each breast: the medio-lateral oblique view (MLO) and the cranio-caudal view (CC) - see Figures 1.1-1.2 for examples of these two views. One of the stages present in the analysis of these mammography views involves the identification and classification of breast lesions, such as breast masses and micro-calcifications (MC) represented by yellow and red

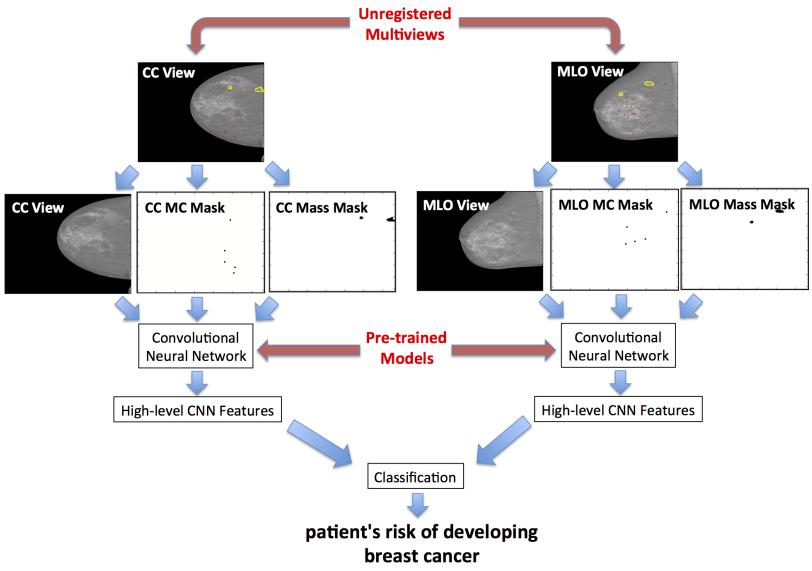

Figure 1.2 The second goal of the paper is to demonstrate how high-level deep learning model features can be used in multi-view mammogram classification problems without pre-registering the mammograms. The meaning of the yellow and red annotations are the same as in Fig. 1.1.

annotations, respectively, in Figures 1.1 and 1.2. This identification and classification is usually performed manually by a radiologist, and a recent assessment of this manual analysis indicates a sensitivity of 84% and a specificity of 91% in the classification of breast cancer [5]. These figures can be improved with the analysis of the mammography views by a second reader: either a radiologist or a computer-aided diagnosis (CAD) system [5]. Therefore, the use of CAD systems as second readers can have a significant impact in breast screening mammography.

Current state-of-the-art CAD systems that can analyse a mammography exam work in delimited stages [5, 6, 7]: lesion detection, lesion segmentation, and lesion classification. The main challenges faced by such systems are related to the low signal-to-noise ratio present in the imaging of the lesion, the lack of a consistent location, shape and appearance of lesions, and the analysis of each lesion independently of other lesions or the whole mammogram. The detection of lesions usually follow a two-step process that first identifies a large number of lesion candidates that are then selected with the goal of reducing false positives while keeping the true positives [8, 9, 10, 11, 12, 13, 14, 15]. Lesion segmentation methods are generally based on global/local energy minimisation models that work on a continuous or discrete space [16, 17, 18]. The final stage consists of the classification of the segmented le-

sions based on typical machine learning classifiers that use as input hand-crafted features extracted from the image region containing the detected lesion and the respective segmentation map [19, 20, 21]. The state-of-the-art binary classification of breast lesions into benign or malignant [22, 23] produces an area under the receiver operating characteristic (ROC) curve between $[0.9, 0.95]$. More similar to our approach, the multi-modal analysis that takes lesions imaged from several modalities (e.g., mammograms and sonograms) have been shown to improve the average performance of radiologists [24]. Note that these hand-crafted features do not guarantee optimality with respect to the classification goal, and the isolated processing of each lesion without looking at the other parts of the exam may ignore crucial contextual information that could help the classification of the whole exam.

In this chapter, we propose a methodology that can analyse a mammography exam in a holistic manner. Specifically, we introduce the design and implementation of a deep learning model [25, 26] that takes as input the two mammography views (CC and MLO) and all detected breast masses and micro-calcifications, and produce an output consisting of a three-class classification of the whole exam: negative (or normal), benign or malignant findings. The challenges present in the development of such deep learning model are: 1) the high capacity of such model can only be robustly estimated with the use of large annotated training sets, and 2) the holistic analysis may require the CC and MLO views to be registered in order to allow the alignment of lesions between these two views. Given that publicly available mammogram datasets do not contain enough annotated samples to robustly train deep learning models, we propose the use of transfer learning [27], where a model is pre-trained with a large annotated computer vision dataset [2], containing typical pictures taken from digital cameras, and fine-tuned with the relatively smaller mammogram datasets. Furthermore, the registration of the CC and MLO views of a mammography exam is a challenging task given the difficulty in estimating the non-rigid deformations that can align these two views, so we propose the classification from the deep learning features, where the hypothesis is that the high-level nature of these features will reduce the need for a low-level matching of the input data [28]. Finally, compared to the previous state of the art in the field, deep learning model can extract features that are automatically learned (as opposed to the previously proposed hand-crafted features) using objective functions formulated based on the classification problem. We test our approach on two publicly available datasets (InBreast [29] and DDSM [30]), and results show that our approach produces a volume under ROC surface of over 0.9 and an area under ROC curve (for a two-class problem: benign and malignant) of over 0.9. The results provide evidence that our method can produce state-of-the-art classification results using a new holistic way of addressing medical image analysis problems.

2. Literature Review

Deep learning has been one of the most studied topics in the fields of computer vision and machine learning [25] for at least three decades. The recent availability of large annotated training sets [2] combined with a competent use of graphics processing units (allowing fast training processes) has enabled the development of classification systems that are significantly more accurate than more traditional machine learning methods [26, 31, 32, 33]. The impact in medical image analysis has been relatively smaller, but also significant [34, 35]. Deep learning has several advantages, compared with traditional machine learning methods [36], such as: features of different abstraction levels are automatically learned using high-level classification objective functions [28]; and methodologies can be designed "end-to-end", where the system can learn how to extract image features, detect and segment visual objects of interest and classify the scene using a unified classification model [37]. The major challenge present in deep learning models is the extremely high number of parameters to estimate during the training process, which requires an equally large annotated training set to enable a robust training process. This challenge is particularly critical in medical image analysis (MIA) applications due to the limited availability of large annotated training set. In fact, the largest MIA datasets typically have in the order of a few thousands of samples, which is generally considered to be not enough for a robust training of a deep learning model. The initial successful MIA applications have been achieved exactly with problems that contain large annotated training sets, such as the mitosis detection [34] and lymph node detection [35]. However, MIA problems that have limited training sets have been generally tackled with the help of regularisation methods, such as unsupervised training [38, 39, 40, 41]. The use of registered multi-view input data has also been tested with deep auto-encoders [42, 43], which is similar to our methodology, except that our multi-view data is not aligned.

Recently, there has been considerable interest in the development of deep learning methods for the analysis of mammograms, where this analysis can be divided into three stages [5]: 1) detection of lesions (i.e., masses and micro-calcifications), 2) segmentation of the detected lesions from the first stage, and 3) classification of the lesions based on texture and shape features extracted from the segmented lesions. The problem of mass detection has been traditionally addressed by classical image processing methods for initial candidate generation, followed by a cascade of machine learning techniques to eliminate false positives [5]. The use of a cascade of deep learning models for mass detection essentially follows the same approach, with the exception that it does not rely on classical image processing methods to generate initial candidates [44]. The use of deep learning methods for lesion segmentation has been explored in different ways. For instance, a straightforward deep learning model receives the image at the input and produces a binary segmentation map at the output [45, 46], which only

works if the annotated training set is relatively large [30]. When the annotated training set is small [21], Dhungel et al. [47, 48, 49] have proposed a method that combines a conditional random field with several deep learning potential functions, where the idea is that this combination will compensate for the small annotated sets used in the training process. Finally, the last stage in the analysis, i.e., lesion classification, has also been addressed with deep learning methods, with a direct classification of the detected and segmented lesions from the first and second stages of the analysis [50, 51, 52].

Deep learning is also allowing the development of methods that can analyse mammograms in a more holistic manner, like the work proposed in this chapter, which represents a clear divergence from the traditional 3-stage analysis process [5] mentioned above. For example, the risk of developing breast cancer can be assessed with deep learning classifiers that score breast density and texture [53, 54]. Finally, Qiu et al. [55] propose a method that estimates the risk of developing breast cancer from a normal mammogram. We expect that such deep learning-based methods that receive a mammography exam at the input and produce either a diagnosis or prognosis result will become the mainstream of future methodologies.

3. Methodology

For the training and testing of our methodology, we have the following dataset: $\mathcal{D} = \{(\mathbf{x}^{(p,b)}, \mathbf{c}^{(p,b)}, \mathbf{m}^{(p,b)}, \mathbf{y}^{(p,b)})\}_{p\in\{1,...,P\},b\in\{\text{left,right}\}}$, with $\mathbf{x} = \{\mathbf{x}_{\text{CC}}, \mathbf{x}_{\text{MLO}}\}$ denoting the mammography views CC and MLO, where $\mathbf{x}_{\text{CC}}, \mathbf{x}_{\text{MLO}} : \Omega \rightarrow \mathbb{R}$ and Ω denotes the image lattice, $\mathbf{c} = \{\mathbf{c}_{\text{CC}}, \mathbf{c}_{\text{MLO}}\}$ representing the micro-calcifications (MC) segmentation in each view with $\mathbf{c}_{\text{CC}}, \mathbf{c}_{\text{MLO}} : \Omega \rightarrow \{0, 1\}$, $\mathbf{m} = \{\mathbf{m}_{\text{CC}}, \mathbf{m}_{\text{MLO}}\}$ denoting the mass segmentation in each view with $\mathbf{m}_{\text{CC}}, \mathbf{m}_{\text{MLO}} : \Omega \rightarrow \{0, 1\}$, $\mathbf{y} \in \mathcal{Y} = \{0, 1\}^C$ being the BI-RADS classification with C classes, $p \in \{1, ..., P\}$ indexing the patients, and $b \in \{\text{left, right}\}$ indexing the patient's left and right breasts (each patient's breast is denoted as a case because they can have different BI-RADS scores). There are six possible BI-RADS classes: 1: negative, 2: benign finding(s), 3: probably benign, 4: suspicious abnormality, 5: highly suggestive of malignancy, 6: proven malignancy. However, the datasets available for this research only contains limited amounts of training data per class, as shown in Fig. 1.6, so we propose the following three-class division of the original classes: negative, denoted by $\mathbf{y} = [1, 0, 0]^\top$, when BI-RADS=1; benign, represented by $\mathbf{y} = [0, 1, 0]^\top$, with BI-RADS $\in \{2, 3\}$; and malignant, denoted by $\mathbf{y} = [0, 0, 1]^\top$, when BI-RADS $\in \{4, 5, 6\}$. The dataset of non-mammography images, used for pre-training the deep learning model, is represented by $\widetilde{\mathcal{D}} = \{(\widetilde{\mathbf{x}}^{(n)}, \widetilde{\mathbf{y}}^{(n)})\}_n$, with $\widetilde{\mathbf{x}} : \Omega \rightarrow \mathbb{R}$ and $\widetilde{\mathbf{y}} \in \widetilde{\mathcal{Y}} = \{0, 1\}^{\widetilde{C}}$, where $\widetilde{C}$ represents the cardinality of the set of classes in the dataset $\widetilde{\mathcal{D}}$. Fig. 1.3 shows examples of the non-mamographic images in $\widetilde{\mathcal{D}}$, and also the mammography views plus their respective binary MC and mass segmentation masks in $\mathcal{D}$.

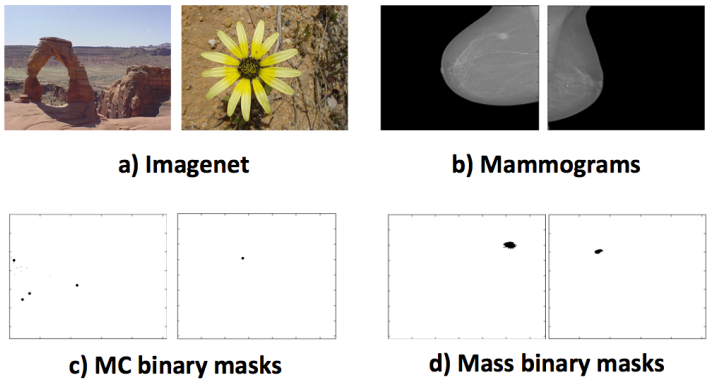

Figure 1.3 The images in (a) represent samples from the dataset $\widetilde{\mathcal{D}}$, used for pre-training the deep learning model, while (b)-(d) display training images and binary maps of micro-calcifications (MC) and masses from dataset $\mathcal{D}$.

3.1. Deep Learning Model

The deep learning model explored in this work consists of the convolutional neural network (CNN), which is represented by $f : \mathcal{X} \rightarrow \mathcal{Y}$ ($\mathcal{X}$ denotes the image or binary segmentation map spaces while $\mathcal{Y}$ represents the space of classification vectors):

$$f(\mathbf{x}, \theta) = f_{\text{out}} \circ f_{fc} \circ h_L \circ g_L \circ f_L \circ ... \circ h_1 \circ g_1 \circ f_1(\mathbf{x}), \tag{1.1}$$

where $\circ$ represents the composition operator, $\{f_i(.)\}_{i=1}^{L}$ denotes the convolutional layers, θ represents the model parameters formed by the input weight matrices $\mathbf{W}_l \in \mathbb{R}^{k_l \times k_l \times n_l \times n_{l-1}}$ and bias vectors $\mathbf{b}_l \in \mathbb{R}^{n_l}$ for each layer $l \in \{1, ..., L\}$, with $k_l \times k_l$ representing the filter size of the n_l filters in layer l that has n_{l-1} input channels, $g_l(.)$ is a non-linear activation layer, $h_l(.)$ is a sub-sampling layer, f_{fc} denotes the set of fully-connected layers $\{\mathbf{W}_{fc,k}\}_{k=1}^{K}$ (with $\mathbf{W}_{fc,k} \in \mathbb{R}^{n_{fc,k-1} \times n_{fc,k}}$ representing the connections from fully connected layer $k-1$ to k) and biases $\{\mathbf{b}_{fc,k}\}_{k=1}^{K}$ (with $\mathbf{b} \in \mathbb{R}^{n_{fc,k}}$) that are also part of the model parameters θ, and f_{out} is a multinomial logistic regression layer [26] that contains weights $\mathbf{W}_{out} \in \mathbb{R}^{n_{fc,K} \times C}$ and bias $\mathbf{b}_{out} \in \mathbb{R}^{C}$, which also belong to θ (Fig. 1.4 shows a visual description of this model). The convolutional layer is defined by

$$\mathbf{F}_l = f_l(\mathbf{x}_{l-1}) = \mathbf{W}_l \star \mathbf{X}_{l-1}, \tag{1.2}$$

where the bias term $\mathbf{b}_l$ is excluded to simplify the equation and we are abusing the notation by representing the convolution of n_{l-1} channels of input $\mathbf{X}_{l-1} = [\mathbf{x}_{l-1,1}, ..., \mathbf{x}_{l-1,n_{l-1}}]$ with the n_l filters of matrix $\mathbf{W}_l$, with $\star$ denoting the convolution operator. The input $\mathbf{X}_{l-1}$ of (1.2) is obtained from the activation (e.g., logistic or rectified linear [26]) and

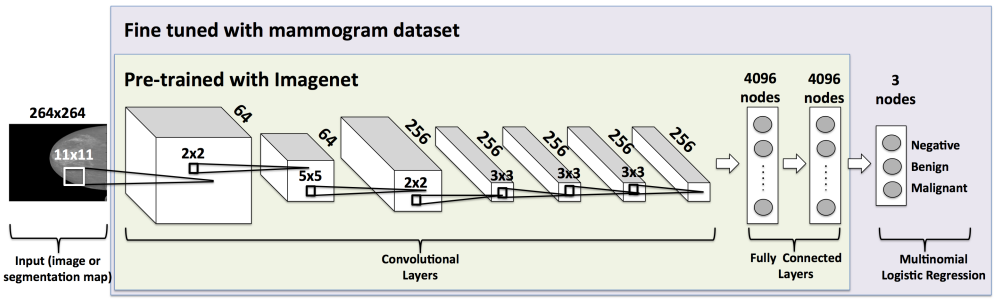

Figure 1.4 Visualisation of the single view CNN model used in this work, containing $L = 5$ stages of convolutional layers, $K = 2$ stages of fully connected layers and one final layer containing the softmax layer.

sub-sampling (e.g., the mean or max pooling functions [26])) of the preceding layer by $\mathbf{X}_{l-1} = h_{l-1}(g_{l-1}(f_{l-1}(\mathbf{X}_{l-2})))$, where $\mathbf{X}_0$ represents the input mammogram $\mathbf{x}$ or segmentation maps $\mathbf{c}$ or $\mathbf{m}$. The output from (1.2) is $\mathbf{F}_l = [\mathbf{f}_{l,1}, ..., \mathbf{f}_{l,n_l}]$, which is a volume containing n_l pre-activation matrices. The L convolutional layers are followed by a sequence of fully connected layers that vectorise the input volume $\mathbf{X}_L$ into $\mathbf{x}_L \in \mathbb{R}^{|\mathbf{x}_L|}$ (where $|\mathbf{x}_L|$ denotes the length of the vector $\mathbf{x}_L$) and apply a couple of linear transforms [26]:

$$\mathbf{f}_{fc} = f_{fc}(\mathbf{X}_L) = \left(\mathbf{W}_{fc,2}\left(\mathbf{W}_{fc,1}\mathbf{x}_L + \mathbf{b}_{fc,1}\right) + \mathbf{b}_{fc,2}\right), \tag{1.3}$$

where the output is a vector $\mathbf{f}_{fc} \in \mathbb{R}^{n_{fc,2}}$. Finally, these fully connected layers are followed by a classification layer defined by a softmax function over a linearly transformed input, as follows [26] :

$$\mathbf{f}_{out} = f_{out}(\mathbf{f}_{fc}) = softmax(\mathbf{W}_{out}\mathbf{f}_{fc} + \mathbf{b}_{out}), \tag{1.4}$$

where $softmax(\mathbf{z}) = \frac{e^{\mathbf{z}}}{\sum_j e^{\mathbf{z}(j)}}$, and $\mathbf{f}_{out} \in [0,1]^C$ represent the output from the inference process that takes $\mathbf{x}$ as the input (recall that the input can be either a mammogram or a segmentation map of a micro-calcification or a mass), with C representing the number of output classes.

Finally, estimating θ in (1.1) involves a training process that is carried out with stochastic gradient descent to minimise the cross entropy loss [26] over the training set, as follows [26]:

$$\ell(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i \log \mathbf{f}_{out,i}^{\top}, \tag{1.5}$$

where N denotes the number of cases available for training, which are indexed by i.

The training of the model in (1.1) comprises two steps: a pre-training stage us-

ing the dataset of non-mammography images $\widetilde{\mathcal{D}}$ and a fine-tuning stage that relies on the dataset of mammography images and segmentation maps $\mathcal{D}$. The pre-training process produces the model $\widetilde{\mathbf{y}}^* = f(\widetilde{\mathbf{x}}; \widetilde{\theta})$, where $\widetilde{\theta} = [\widetilde{\theta}_1, ..., \widetilde{\theta}_L, \widetilde{\theta}_{fc,1}, \widetilde{\theta}_{fc,2}, \widetilde{\theta}_{out}]$ with $\widetilde{\theta}_l = [\widetilde{\mathbf{W}}_l, \widetilde{\mathbf{b}}_l]$ denoting the parameters of the convolutional layer $l \in \{1, ..., L\}$, $\widetilde{\theta}_{fc,k} = [\widetilde{\mathbf{W}}_{fc,k}, \widetilde{\mathbf{b}}_{fc,k}]$ representing the parameters of the fully connected layer $k \in \{1, 2\}$, and $\widetilde{\theta}_{out} = [\widetilde{\mathbf{W}}_{out}, \widetilde{\mathbf{b}}_{out}]$ denoting the parameters of the softmax layer. This pre-training is carried out by minimising the cross-entropy loss in (1.5) with the $\widetilde{C}$ classes present in the dataset $\widetilde{\mathcal{D}}$. The fine-tuning process takes a subset of $\widetilde{\theta}$ comprising $[\widetilde{\theta}_1, ..., \widetilde{\theta}_L, \widetilde{\theta}_{fc,1}, \widetilde{\theta}_{fc,2}]$ (i.e., all parameters except for $\widetilde{\theta}_{out}$) to initialise the new training parameters $\theta = [\theta_1, ..., \theta_L, \theta_{fc,1}, \theta_{fc,2}, \theta_{out}]$, where θ_{out} is initialised with random values, and all parameters in θ are re-trained to minimise the cross-entropy loss in (1.5) with the C classes in $\mathcal{D}$ (see Fig. 1.4). Recently published results [27] have shown that such fine-tuning process depends on the use of a large number of pre-trained layers, which explains why we initialise almost all parameters (except for θ_{out}) with the values estimated from the pre-training process. This fine-tuning shall produce six models: 1) MLO image model: $\mathbf{y} = f(\mathbf{x}_{\text{MLO}}; \theta_{\text{MLO,im}})$, 2) CC image model $\mathbf{y} = f(\mathbf{x}_{\text{CC}}; \theta_{\text{CC,im}})$, 3) MLO MC segmentation map model $\mathbf{y} = f(\mathbf{c}_{\text{MLO}}; \theta_{\text{MLO,mc}})$, 4) CC MC segmentation map model $\mathbf{y} = f(\mathbf{c}_{\text{CC}}; \theta_{\text{CC,mc}})$, 5) MLO mass segmentation map model $\mathbf{y} = f(\mathbf{m}_{\text{MLO}}; \theta_{\text{MLO,ma}})$ and 6) CC mass segmentation map model $\mathbf{y} = f(\mathbf{m}_{\text{CC}}; \theta_{\text{CC,ma}})$.

Finally, the multi-view model combines the six models produced from the fine-tuning process by concatenating the features from the last fully connected layer of all six models, represented by $[\mathbf{f}_{fc,i}]_{i \in \{\text{MLO,im,CC,MLO,im,MLO, mc,CC,mc,MLO,ma,CC,ma}\}}$, and training a single multinomial logistic regression layer using those inputs (Fig. 1.5). This multi-view model is represented by:

$$\mathbf{f}_{out,mv} = softmax(\mathbf{W}_{out,mv}[\mathbf{f}_{fc,i}]_{i \in \{\text{MLO,im,CC,MLO,im,MLO, mc,CC,mc,MLO,ma,CC,ma}\}} + \mathbf{b}_{out,mv}), \tag{1.6}$$

and trained by minimising the cross-entropy loss in (1.5) with the C classes in $\mathcal{D}$, where $\theta_{mv} = [\mathbf{W}_{out,mv}, \mathbf{b}_{out,mv}]$ is randomly initialized in this multi-view training.

4. Materials and Methods

For the experiments below, we use two mammogram datasets that are publicly available: InBreast [29] and DDSM [30]. The InBreast [29] dataset contains 115 patients, where there are around four images per patients, amounting to 410 images. InBreast does not come with a suggested division of training and testing sets, so our experimental results are based on a five-fold cross validation, where each fold uses a division of 90 patients for training and 25 patients for testing. The DDSM [30] dataset contains 172 patients, each having around four images, which results in 680 images. This dataset is formed by merging the original micro-calcification and mass datasets,

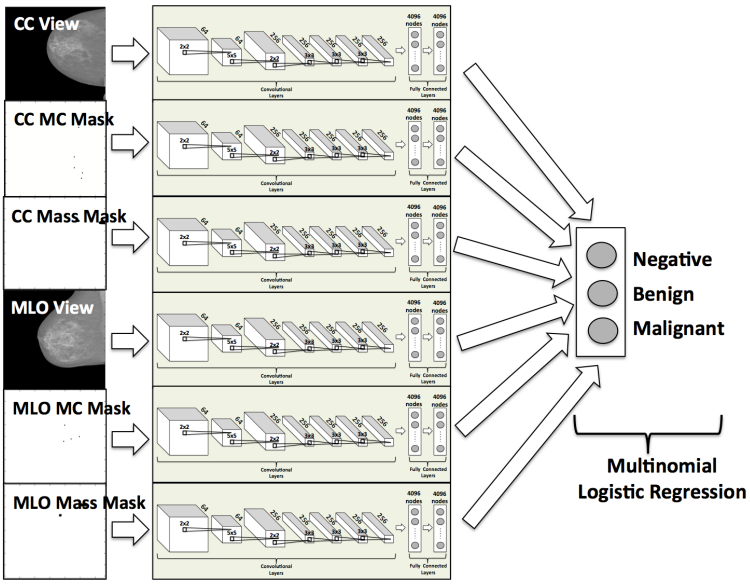

Figure 1.5 Visualisation of the multi-view model with the responses from the single view CNN models (see Fig. 1.4) that are connected to a classification layer.

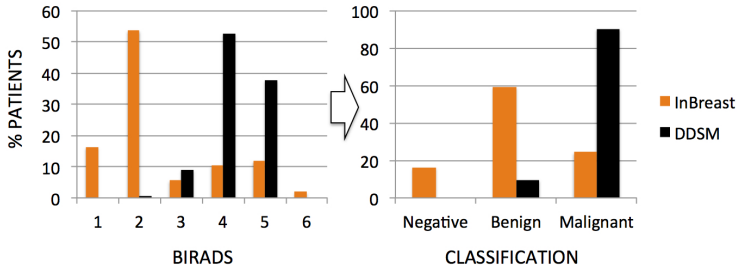

Figure 1.6 Distribution of BI-RADS (left) and negative, benign and malignant classes (right) for the cases in InBreast (orange) and DDSM (black).

but removing the overlapping cases that are available from the training set of mass and testing set of micro-calcification and vice-versa. We use the suggested division of training and testing sets for DDSM [30], containing 86 patients for training and 86 for testing. It is important to notice that the distributions of BI-RADS and, consequently the negative, benign and malignant classes in InBreast and DDSM are quite different, as shown in Fig. 1.6. In particular, InBreast tries to keep the percentage of negative (i.e., normal) and benign cases at a higher level than the malignant cases, while DDSM has a much larger percentage of malignant cases, compared to benign

and negative cases.

The CC and MLO mammography views are pre-processed with local contrast normalisation, which is followed by Otsu's segmentation [56] that crops out the image region containing the background. The remaining image is scaled (using bi-cubic interpolation) to have size 264×264. Furthermore, a simple algorithm is run in order to flip the mammograms such that the pectoral muscle always lies on the right-hand side of the image. The manual binary segmentation maps representing the micro-calcification and mass present in a mammography view uses the same geometric transformations applied to their respective views (i.e., the cropping, scaling and flipping). If no micro-calcification and mass is present in a particular view, then we use a 264×264 image filled with zeros (i.e., a blank image). Fig. 1.9 shows some samples of the pre-processed mammography views and their respective segmentation maps.

The base CNN model is based on Chatfield et al.'s CNN-F model [57], which is a simplified version of AlexNet [26], containing fewer filters. Figure 1.4 shows the details of the CNN-F model, where the input image has size 264×264, the first convolutional stage as 64 11×11 filters and a max-pooling that sub-samples the input by 2, the second convolutional stage has 256 5×5 filters and a max-pooling that sub-samples the input by 2, the third, fourth and fifth convolutional stages have 256 3×3 filters (each) with no sub-sampling, the first and second fully connected stages have 4096 nodes (each), and the multinomial logistic regression stage contains softmax layer with three nodes. We use the CNN-F model that Chatfield et al. [57] have pre-trained with Imagenet [2] (1K visual classes, 1.2M training, 50K validation and 100K test images). The fine-tuning process consists of replacing the multinomial logistic regression stage at the end by a new layer that has three classes (negative, benign and malignant) and train it for the CC and MLO views, the micro-calcification segmentation maps of the CC and MLO views, and the mass segmentation maps of the CC and MLO views (see Fig. 1.4). The multi-view model is built by concatenating the 4096-dimensional feature vectors available from the second fully connected stages of the six models (forming a 24576-dimensional feature vector) and training a single multinomial logistic regression with three nodes (see Fig. 1.5). This two-stage training, comprising pre-training and fine-tuning, can be seen as a regularisation that helps the generalisation ability of the model. As a result, we can compare such two-stage training to other forms of regularisation, such as data augmentation [26], which is obtained by applying random geometric transformations to the original training images in order to generate new artificial training samples. We compare our proposed two-stage training with data augmentation with an experiment that uses the CNN-F structure defined above without pre-training, which means that the training for the parameter θ in (1.1) is randomly initialised using an unbiased Gaussian with standard deviation 0.001, and run with data augmentation by adding 5, 10 and 20 new samples per training image. In this data augmentation training, each original training image is randomly cropped

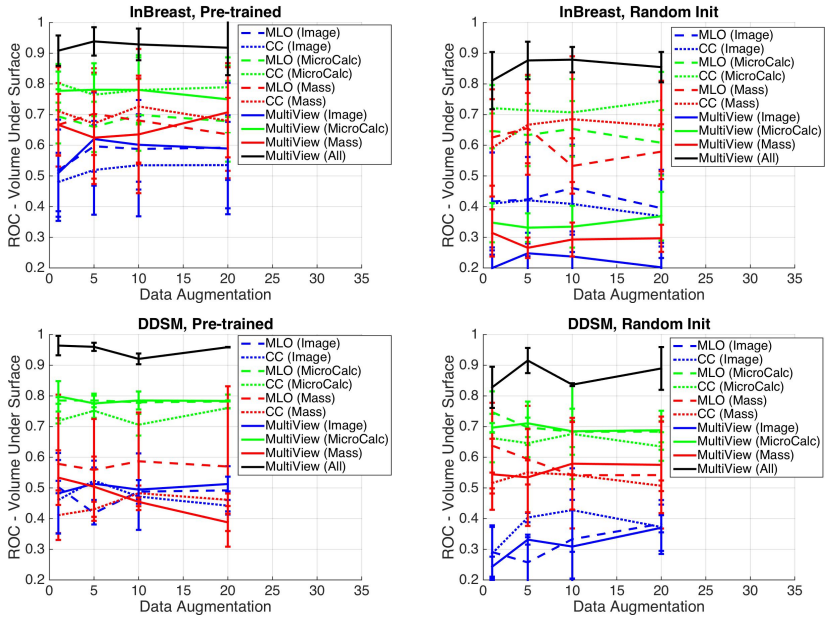

Figure 1.7 VUS in terms of data augmentation on InBreast (top) DDSM (bottom) for the MLO and CC views, and with each isolate input (image, micro-calcification and mass segmentation maps), all inputs (All) and both views (Multiview) together. 1^{st} column shows the results with the Imagenet pre-trained model, and the 2^{nd} column shows the randomly initialized models.

from the top-left and bottom-right corners within a range of $[1, 10]$ pixels from the original corners. This data augmentation is also used in the two-stage training process in order to verify if the combination of two regularisation methods can improve the generalisation of the CNN-F model. In all training processes, the learning rate is fixed at 0.001 and momentum is 0.9.

Classification accuracy is measured in two ways. For a three-class problem, with classes negative, benign and malignant, the accuracy is measured with the volume under ROC surface (VUS) [58]. The two-class problem, with classes benign and malignant, is assessed by the area under ROC curve (AUC), where it is assumed that all cases contain at least one finding (a micro-calcification or a mass).

5. Results

The VUS as a function of the data augmentation (varying in the range $\{0, 5, 10, 20\}$) for the test sets of InBreast and DDSM are shown in Figure 1.7. For InBreast, the results are shown with the average and standard deviation of 5-fold cross validation and the two breasts, and for DDSM, results are based on the average and standard

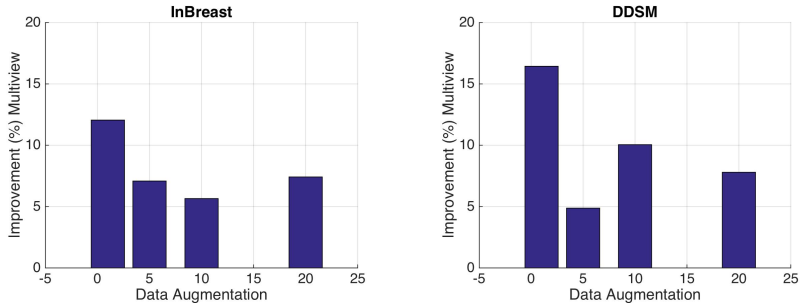

Figure 1.8 Mean improvement of the VUS results for InBreast (left) and DDSM (right) of the pre-trained models compared to the randomly initialised models.

deviation of the two breasts. Also note that in Fig. 1.7 we show the results for the MLO and CC views, and with each isolate input (image and segmentation maps) in addition to all inputs and both views together. The improvement achieved with the use of the Imagenet pre-trained model, compared with the randomly initialised model, is shown in Fig. 1.8. We also show five examples from the classification process using the pre-trained model (without data augmentation) on InBreast test cases in Figure 1.9.

Focusing on the cases where there exists at least one lesion (mass or micro- calcification) allows us to compute the AUC for a two-class problem (benign or malignant). Using the model that is pre-trained with Imagenet and fine-tuned with InBreast without data augmentation, produces an AUC of 0.91(±0.05), and fine-tuned with DDSM results in an AUC of 0.97(±0.03).

Finally, running Matconvnet [57] on a standard desktop (2.3GHz Intel Core i7 with 8GB, and graphics card NVIDIA GeForce GT 650M 1024 MB), the training time for six models and the multiview model (without data augmentation) is one hour. With the addition of 10 artificial training samples per original training sample, the training time increases to four hours, and with 20 artificial training samples per original training sample, the training time increases to seven hours.

6. Discussion

The graphs in Figure 1.7 show that the multiview results that use all inputs (images and segmentation maps), represented by the solid black curve, present the best performance amongst all models considered in this work. This shows evidence that the high-level features provided by each model are indeed useful for the classification of the whole exam, even though the input images and segmentation maps are not registered. Another interesting result shown in Figures 1.7 and 1.8 is the improvement of 5% to 16% observed with the use of Imagenet pre-trained models, particularly when

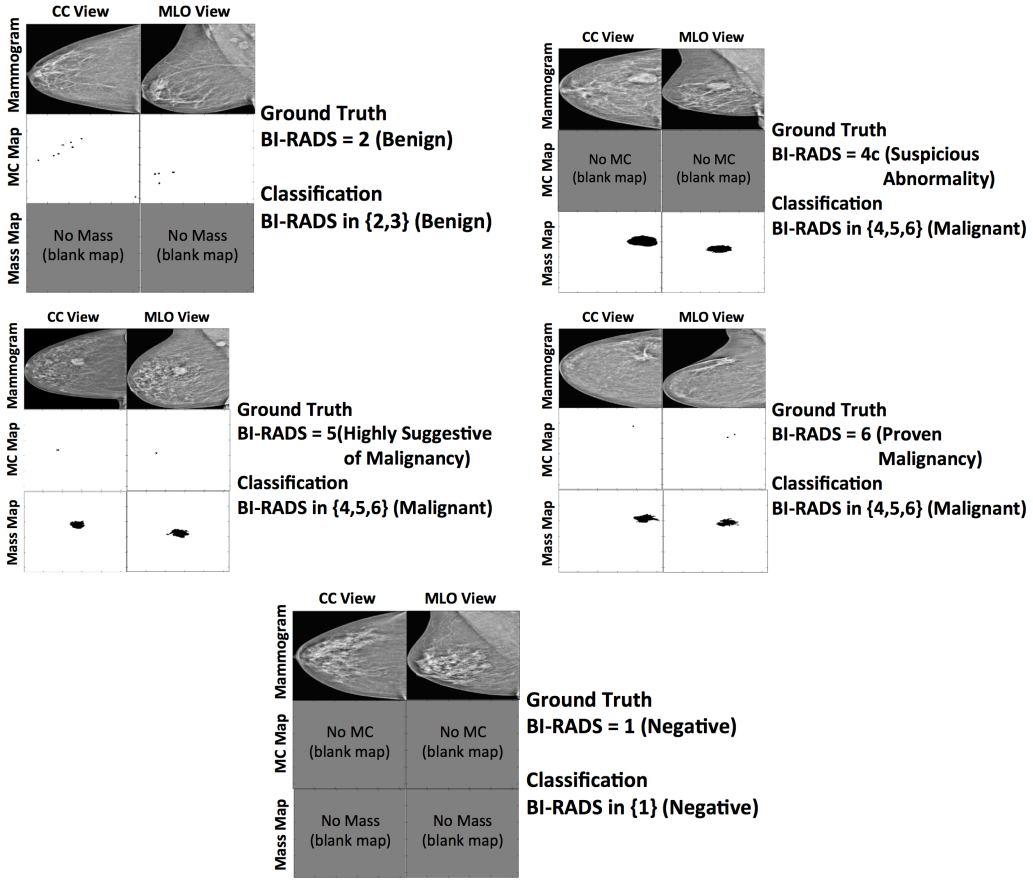

Figure 1.9 InBreast test case results using Imagenet pre-trained model with no data augmentation, where the ground truth and the automatic classifications are shown.

the training process does not involve data augmentation. One final point shown by Fig. 1.7 is that the results, with respect to data augmentation, saturates rather quickly with the use of five or more artificially generated training samples (per each original training sample). This point needs further investigation - for instance, it may the case that geometric transformations may not be the most appropriate way of augmenting medical data. The visual results in Figure 1.9 show that the system is likely to classify cases as malignant when micro-calcifications and masses are found, and as negative when no lesions are found. However, when either masses or micro-calcifications (but not both) are found, then the system can classify the case either as benign or malignant.

The results in Sec. 5 also show poor performance of the single/multi view classi-

fications containing only one of the inputs (image or segmentation maps). This may happen due to several reasons, such as that cases where BI-RADS > 1 may contain annotations for either micro-calcification or mass, but not for both lesions. Also, the mammogram images alone may not have sufficient information for a robust classification, particularly considering the fact that they are down-sampled around ten-fold to an input of size 264×264. It is also interesting to note the consistency of the results in the sense that micro-calcification segmentation maps produce better classification results than mass, which in turn is better than the image classification.

The comparison of our proposed methodology to previously proposed methods in the field (see Sec. 2) is difficult because most of these previous methods use datasets that are not publicly available and they also focus on the classification of individual lesions, as opposed to the classification of the whole exam that we propose. In any case, it is possible to compare the AUC results produced by our method to the AUC results of individual mass/micro-calcification classification of the current state of the art, which are between $[0.9, 0.95]$ for MCs and mass classification [22, 23]. Therefore, we can conclude that our proposed method is comparable (on InBreast) or superior (on DDSM) than the current state of the art.

7. Conclusion

In this chapter, we show that the high level features produced by deep learning models are effective for classification tasks that use un-registered inputs. This is particularly important in mammograms, where registration is challenging due to non-rigid deformations. Moreover, the use of pre-trained models appears to be advantageous, compared to the randomly initialised models. This is somewhat an expected result given that the randomly initialised model is more likely to overfit the training data. We would like to emphasise that the results shown in Sec. 5 can serve as baseline for the field because the data used is publicly available, which allows for a fair comparison with future works that will be published in the field [5]. Our proposal has the potential to open two research fronts that can be applied to other medical image analysis problems: 1) the use of deep learning models pre-trained with non-medical imaging datasets, and 2) the holistic analysis of un-registered multi-view medical images.

ACKNOWLEDGMENTS

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).

s

REFERENCE

[1] Gustavo Carneiro, Jacinto Nascimento, Andrew P Bradley, Unregistered Multiview Mammogram Analysis with Pre-trained Deep Learning Models, in: Medical Image Computing and Computer-Assisted InterventionMICCAI 2015, Springer 2015 pp. 652–660.

[2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei, ImageNet Large Scale Visual Recognition Challenge 2014, arXiv:1409.0575.

[3] Ahmedin Jemal, Rebecca Siegel, Elizabeth Ward, Yongping Hao, Jiaquan Xu, Taylor Murray, Michael J Thun, Cancer statistics, 2008, CA: a cancer journal for clinicians 58 (2) (2008) 71–96.

[4] Béatrice Lauby-Secretan, Chiara Scoccianti, Dana Loomis, Lamia Benbrahim-Tallaa, Véronique Bouvard, Franca Bianchini, Kurt Straif, Breast-cancer screeningviewpoint of the IARC Working Group, New England Journal of Medicine 372 (24) (2015) 2353–2358.

[5] Maryellen L Giger, Nico Karssemeijer, Julia A Schnabel, Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer, Annual review of biomedical engineering 15 (2013) 327–357.

[6] Arnau Oliver, Jordi Freixenet, Joan Marti, Elsa Perez, Josep Pont, Erika RE Denton, Reyer Zwiggelaar, A review of automatic mass detection and segmentation in mammographic images, Medical Image Analysis 14 (2) (2010) 87–110.

[7] Jinshan Tang, Rangaraj M Rangayyan, Jun Xu, Issam El Naqa, Yongyi Yang, Computer-aided detection and diagnosis of breast cancer with mammography: recent advances, Information Technology in Biomedicine, IEEE Transactions on 13 (2) (2009) 236–251.

[8] Ehsan Kozegar, Mohsen Soryani, Behrouz Minaei, Inês Domingues, et al., Assessment of a novel mass detection algorithm in mammograms, Journal of cancer research and therapeutics 9 (4) (2013) 592.

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

[10] Guido M te Brake, Nico Karssemeijer, Jan HCL Hendriks, An automatic method to discriminate malignant masses from normal tissue in digital mammograms, Physics in Medicine and Biology 45 (10) (2000) 2843.

[11] Renato Campanini, Danilo Dongiovanni, Emiro Iampieri, Nico Lanconelli, Matteo Masotti, Giuseppe Palermo, Alessandro Riccardi, Matteo Roffilli, A novel featureless approach to mass detection in digital mammograms based on support vector machines, Physics in Medicine and Biology 49 (6) (2004) 961.

[12] Nevine H Eltonsy, Georgia D Tourassi, Adel Said Elmaghraby, A concentric morphology model for the detection of masses in mammography, Medical Imaging, IEEE Transactions on 26 (6) (2007) 880–889.

[13] Mehul P Sampat, Alan C Bovik, Gary J Whitman, Mia K Markey, A model-based framework for the detection of spiculated masses on mammographya), Medical physics 35 (5) (2008) 2110–2123.

[14] Roberto Bellotti, Francesco De Carlo, Sonia Tangaro, Gianfranco Gargano, Giuseppe Maggipinto, Marcello Castellano, Raffaella Massafra, Donato Cascio, Francesco Fauci, Rosario Magro, et al., A completely automated CAD system for mass detection in a large mammographic database, Medical physics 33 (8) (2006) 3066–3075.

[15] Jun Wei, Berkman Sahiner, Lubomir M Hadjiiski, Heang-Ping Chan, Nicholas Petrick, Mark A Helvie, Marilyn A Roubidoux, Jun Ge, Chuan Zhou, Computer-aided detection of breast masses on full field digital mammograms, Medical physics 32 (9) (2005) 2827–2838.

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

[17] Peyman Rahmati, Andy Adler, Ghassan Hamarneh, Mammography segmentation with maximum likelihood active contours, Medical image analysis 16 (6) (2012) 1167–1186.

[18] Jaime S Cardoso, Inês Domingues, Hélder P Oliveira, Closed shortest path in the original coordinates with an application to breast cancer, International Journal of Pattern Recognition and Artificial Intelligence 29 (01) (2015) 1555002.

© Elsevier Ltd.
All rights reserved.

[19] C Varela, S Timp, N Karssemeijer, Use of border information in the classification of mammographic masses, Physics in Medicine and Biology 51 (2) (2006) 425.
[20] Jiazheng Shi, Berkman Sahiner, Heang-Ping Chan, Jun Ge, Lubomir Hadjiiski, Mark A Helvie, Alexis Nees, Yi-Ta Wu, Jun Wei, Chuan Zhou, et al., Characterization of mammographic masses based on level set segmentation with new image features and patient information, Medical physics 35 (1) (2008) 280–290.
[21] I Domingues, E Sales, JS Cardoso, WCA Pereira, Inbreast-Database masses characterization, XXIII CBEB (2012).
[22] HD Cheng, XJ Shi, Rui Min, LM Hu, XP Cai, HN Du, Approaches for automated detection and classification of masses in mammograms, Pattern recognition 39 (4) (2006) 646–668.
[23] Liyang Wei, Yongyi Yang, Robert M Nishikawa, Yulei Jiang, A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications, Medical Imaging, IEEE Transactions on 24 (3) (2005) 371–380.
[24] Karla Horsch, Maryellen L Giger, Carl J Vyborny, Li Lan, Ellen B Mendelson, R Edward Hendrick, Classification of Breast Lesions with Multimodality Computer-aided Diagnosis: Observer Study Results on an Independent Clinical Data Set, Radiology 240 (2) (2006) 357–368.
[25] Yann LeCun, Yoshua Bengio, Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks 3361 (1995).
[26] Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton, ImageNet Classification with Deep Convolutional Neural Networks., in: NIPS, vol. 1 2012, p. 4.
[27] Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson, How transferable are features in deep neural networks?, in: Advances in Neural Information Processing Systems 2014, pp. 3320–3328.
[28] Yoshua Bengio, Learning deep architectures for AI, Foundations and trends® in Machine Learning 2 (1) (2009) 1–127.
[29] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, Jaime S Cardoso, INbreast: toward a full-field digital mammographic database, Academic Radiology 19 (2) (2012) 236–248.
[30] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, P Kegelmeyer, The digital database for screening mammography, in: Proceedings of the 5th international workshop on digital mammography 2000, pp. 212–218.
[31] Clément Farabet, Camille Couprie, Laurent Najman, Yann LeCun, Learning hierarchical features for scene labeling, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (8) (2013) 1915–1929.
[32] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE 2014, pp. 580–587.
[33] Yuting Zhang, Kihyuk Sohn, Ruben Villegas, Gang Pan, Honglak Lee, Improving Object Detection with Deep Convolutional Networks via Bayesian Optimization and Structured Prediction, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015, pp. 249 – 258.
[34] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Mitosis detection in breast cancer histology images with deep neural networks, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, Springer 2013 pp. 411–418.
[35] Holger R Roth, Le Lu, Ari Seff, Kevin M Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, Ronald M Summers, A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, Springer 2014 pp. 520–527.
[36] Christopher M Bishop, Pattern Recognition, Machine Learning (2006).
[37] Jifeng Dai, Kaiming He, Jian Sun, Instance-aware Semantic Segmentation via Multi-task Network Cascades, arXiv preprint arXiv:1512.04412 (2015).
[38] Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber, Using deep learning to enhance cancer diagnosis and classification, in: Proceedings of the ICML Workshop on the Role of Machine Learning in Transforming Healthcare (WHEALTH). Atlanta, GA 2013.
[39] Gustavo Carneiro, Jacinto C Nascimento, Combining multiple dynamic models and deep learning

architectures for tracking the left ventricle endocardium in ultrasound data, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (11) (2013) 2592–2607.

[40] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, Fabio Augusto González Osorio, A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, Springer 2013 pp. 403–410.

[41] Rongjian Li, Wenlu Zhang, Heung-Il Suk, Li Wang, Jiang Li, Dinggang Shen, Shuiwang Ji, Deep learning based imaging data completion for improved brain disease diagnosis, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, Springer 2014 pp. 305–312.

[42] Tom Brosch, Youngjin Yoo, David KB Li, Anthony Traboulsee, Roger Tam, Modeling the Variability in Brain Morphology and Lesion Distribution in Multiple Sclerosis by Deep Learning, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, Springer 2014 pp. 462–469.

[43] Yanrong Guo, Guorong Wu, Leah A Commander, Stephanie Szary, Valerie Jewells, Weili Lin, Dinggang Shen, Segmenting Hippocampus from Infant Brains by Sparse Patch Matching with Deep-Learned Features, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, Springer 2014 pp. 308–315.

[44] N. Dhungel, G. Carneiro, A.P. Bradley, Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests, in: Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on 2015, pp. 1–8, doi:10.1109/DICTA.2015.7371234.

[45] Anastasia Dubrovina, Pavel Kisilev, Boris Ginsburg, Sharbell Hashoul, Ron Kimmel, Computational Mammography using Deep Neural Networks, in: Workshop on Deep Learning in Medical Image Analysis (DLMIA) 2016.

[46] Mehmet Gunhan Ertosun, Daniel L Rubin, Probabilistic visual search for masses within mammography images using deep learning, in: Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on, IEEE 2015, pp. 1310–1315.

[47] Neeraj Dhungel, Gustavo Carneiro, Andrew P Bradley, Deep Learning and Structured Prediction for the Segmentation of Mass in Mammograms, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Springer 2015 pp. 605–612.

[48] N. Dhungel, G. Carneiro, A. P. Bradley, Tree RE-weighted belief propagation using deep learning potentials for mass segmentation from mammograms, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), ISSN 1945-7928 2015, pp. 760–763, doi:10.1109/ISBI.2015.7163983.

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

[50] John Arevalo, Fabio A González, Raúl Ramos-Pollán, Jose L Oliveira, Miguel Angel Guevara Lopez, Representation learning for mammography mass lesion classification with convolutional neural networks, Computer Methods and Programs in Biomedicine (2016).

[51] Yuchen Qiu, Shiju Yan, Maxine Tan, Samuel Cheng, Hong Liu, Bin Zheng, Computer-aided classification of mammographic masses using the deep learning technology: a preliminary study, in: SPIE Medical Imaging, International Society for Optics and Photonics 2016, pp. 978520–978520.

[52] Zhicheng Jiao, Xinbo Gao, Ying Wang, Jie Li, A deep feature based framework for breast masses classification, Neurocomputing (2016).

[53] Michiel Kallenberg, Kersten Petersen, Mads Nielsen, Andrew Ng, Pengfei Diao, Christian Igel, Celine Vachon, Katharina Holland, Nico Karssemeijer, Martin Lillholm, Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring (2016).

[54] Kersten Petersen, Mads Nielsen, Pengfei Diao, Nico Karssemeijer, Martin Lillholm, Breast tissue segmentation and mammographic risk scoring using deep learning, in: Breast Imaging, Springer 2014 pp. 88–94.

[55] Yuchen Qiu, Yunzhi Wang, Shiju Yan, Maxine Tan, Samuel Cheng, Hong Liu, Bin Zheng, An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology, in: SPIE Medical Imaging, International Society for Optics and Photonics 2016, pp. 978521–978521.

[56] Nobuyuki Otsu, A threshold selection method from gray-level histograms, Automatica 11 (285-296) (1975) 23–27.

[57] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, Andrew Zisserman, Return of the devil in the details: Delving deep into convolutional nets, arXiv preprint arXiv:1405.3531 (2014).

[58] Thomas CW Landgrebe, Robert PW Duin, Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis, Pattern Analysis and Machine Intelligence, IEEE Transactions on 30 (5) (2008) 810–822.