Automated Analysis of Unregistered Multi-view Mammograms with Deep Learning

Gustavo Carneiro, Jacinto Nascimento, Andrew P. Bradley

***Abstract*—We describe an automated methodology for the analysis of unregistered cranio-caudal (CC) and medio-lateral oblique (MLO) mammography views in order to estimate the patient's risk of developing breast cancer. The main innovation behind this methodology lies in the use of deep learning models for the problem of jointly classifying unregistered mammogram views and respective segmentation maps of breast lesions (i.e., masses and micro-calcifications). This is a holistic methodology that can classify a whole mammographic exam, containing the CC and MLO views and the segmentation maps, as opposed to the classification of individual lesions, which is the dominant approach in the field. We also demonstrate that the proposed system is capable of using the segmentation maps generated by automated mass and micro-calcification detection systems, and still producing accurate results. The semi-automated approach (using manually defined mass and micro-calcification segmentation maps) is tested on two publicly available datasets (INbreast and DDSM), and results show that the volume under ROC surface (VUS) for a 3-class problem (normal tissue, benign and malignant) is over 0.9, the area under ROC curve (AUC) for the 2-class "benign vs malignant" problem is over 0.9, and for the 2-class breast screening problem (malignancy vs normal/benign) is also over 0.9. For the fully automated approach, the VUS results on INbreast is over 0.7, and the AUC for the 2-class "benign vs malignant" problem is over 0.78, and the AUC for the 2-class breast screening is 0.86.**

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

I. Introduction

Recently published data suggests that breast cancer is responsible for 23% of all cancer cases and 14% of cancer related deaths amongst women worldwide [45]. One of the most effective tools in the reduction of morbidity and mortality associated with breast cancer is based on its early detection via the analysis of two mammographic views of each breast [52]: the medio-lateral oblique view (MLO) and the cranio-caudal view (CC) - see Fig. 1. This analysis is essentially based on the detection and classification of breast lesions (note the yellow and red contours of breast masses and micro-calcifications (MC) in Fig. 1), which is usually manually performed by a radiologist - a recent study indicates that this manual analysis has a specificity of 91% and a sensitivity of 84% in the classification of breast cancer [38]. Giger et al. [38] have suggested that such performance can be improved with the use of a second reading of the same mammogram either by another radiologist or by a computer-aided diagnosis (CAD) system [38]. Hence, the development of CAD systems that can be used as adjunct reader is an important step towards the acceptance of such systems in clinical practice.

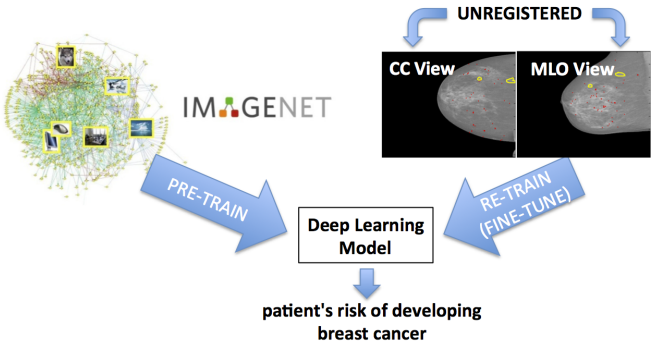


Fig. 1. The main contribution of this paper is the joint analysis of the unregistered cranio-caudal (CC) and medio-lateral oblique (MLO) mammography views with the automatically generated mass (yellow annotations) and micro-calcification (red annotations) segmentation maps. This is a holistic methodology that can classify a whole mammographic exam, with the CC and MLO views and the segmentation maps, as opposed to the classification of individual lesions, which is the mainstream approach of the field. The functionality of our methodology relies on the use of deep learning models, pre-trained with computer vision datasets [4,5,15,81].

The vast majority of mammogram analysis systems are focused on the analysis (i.e., detection, segmentation and classification) of individual lesions (e.g., masses or MCs) [38,59,75] using hand-crafted image features and traditional machine learning classifiers [9]. The outcome of this analysis usually consists of the classification of each lesion into benign or malignant. Lesion detection methods are usually based on a cascade of classifiers that aim to eliminate an increasingly larger number of false positives while keeping a large proportion of the true positives [6,7,12,22,30,48,49,67,76,78]. The assumption that not only the appearance, but also the shape of a lesion is important in its classification motivates the development of lesion segmentation methods [3,13,23,64]. The final lesion classification step generally uses hand-crafted appearance and shape features, extracted from the detected and segmented lesion, that are used as input to a binary classifier that classifies the lesion into benign or malignant [19,28,71, 77,79]. The use of multiple views of the same lesion has

G. Carneiro is with the Australian Centre for Visual Technologies, University of Adelaide, Australia; J. Nascimento is with the Institute for Systems and Robotics, Instituto Superior Técnico, Portugal; and A. Bradley is with the School of Information Technology and Electrical Engineering, University of Queensland, Australia.

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).

also been explored [38,43], and current evidence suggests that such approaches can potentially improve the performance of the system. The main issues with these approaches lie in the sub-optimality (with respect to the classification goal) inherent to the process of hand-crafting features (a notable exception is the lesion detection method by Kooi et al. [48]), and the independent analysis of each lesion that ignores dependencies and contextual information.

In this paper, we propose a new methodology that analyses a two-view mammographic exam in a fully automated and holistic manner. The main innovation behind our approach is the use of a deep learning model [50,53] that receives as input, both the CC and MLO mammographic views and the segmentation maps of the breast lesions (i.e., masses and MCs) and outputs a classification of the exam into normal tissue, benign or malignant (hereafter, we refer to the normal tissue class as negative). The proposed methodology faces the following challenges: 1) deep learning models need annotated datasets that are orders of magnitude larger than what is currently available in medical imaging, and 2) the joint analysis of unregistered multi-view (CC and MLO) and multi-modal inputs (images and segmentation maps) require high-level features that represent the global information present in those inputs.

The first challenge is addressed with transfer learning [4, 5,81], where the deep learning model is first trained with a large annotated computer vision dataset [66], and then re-trained (or fine-tuned) using small annotated mammogram datasets. In parallel to the development of our own work, other similar approaches have been proposed, such as the use of ImageNet [66] to pre-train a deep learning model that identifies pathologies in chest x-ray images [4,5], or the thorough study produced by Tajbakhsh et al. [74] on the use of non-medical image datasets to pre-train deep learning models to be used in various medical image analysis tasks. The second challenge is solved with the use of the high-level features produced by the deep learning models, where we assume that the high-level nature of the deep learning features reduces the need for a low-level matching (registration) of the input data [8]. After the development of our original work [15], which is extended in this paper, there have been relatively similar proposals that classify whole or large patches of mammograms using deep learning models [27,37,44]. In fact, Dhungel et al.'s and Geras et al.'s [27,37] approaches represent extensions to our own original work [15]. We test two versions of our proposed methodology: a semi-automated approach that uses the manually defined segmentation maps of the lesions, and a fully automated approach that uses the lesion detection results from Dhungel et al. [22] and Lu et al. [55]. Compared to previously proposed methods in the field, our model is able to automatically learn the features that are optimal for the classification problem (as opposed to hand-crafting them) and to process a full mammographic exam in a holistic manner, without making lesion independence assumptions. The semi-automated approach is assessed on two publicly available datasets (INbreast [57] and DDSM [42]), where it produces state of the art results for the 3-class and 2-class classification problems. The fully automated approach assessed on INbreast [57] shows a competitive result with respect to the semi-automated approach on the same classification problems.

This paper is an extension of two preliminary works [14,15], where the innovations consist of: 1) the fully automated methodology based on automatically detected masses and MCs, 2) a study on the stage of the deep learning model to merge the different modalities, 3) a study involving a larger set of data augmentation, and 4) a new way of joining the input images as 3-D inputs rather than a collection of 2-D data [1].

II. Literature Review

Deep learning models have been studied for decades [53], but only recently they have achieved important breakthroughs in computer vision and machine learning [34,39,50,82]. This achievement can be explained by the availability of large annotated training sets [66] and the fast training allowed by graphics processing units. Compared to traditional machine learning models [9], deep learning models offer the opportunity to automatically learn features of different abstraction levels directly from raw input, based on high-level classification objective functions [8] and can facilitate the use of multi-modal inputs [58]. There are currently four main trends in the development of deep learning models for medical image analysis. The first is on the acquisition of massive training sets containing only the original annotations that are already present in the dataset (e.g., diagnosis, radiology reports, and not manual delineations of lesions). These methods are producing outstanding results, which are comparable to expert radiologists' performance, e.g., Esteva et al. [32] have developed a deep learning model capable of classifying skin lesions trained directly from image pixels and disease labels as inputs, using a dataset of 129,450 clinical images. This model achieves competitive performance with respect to 21 board-certified dermatologists - unprecedented in terms of the scale of the training set and the accuracy of the classification. Similarly, a recently deep learning method developed by Gulshan et al. [40] has shown to have high sensitivity and specificity for detecting referable diabetic retinopathy in retinal fundus photographs, where the training set contained 128,175 annotated retinal images. A similar work has been proposed by Geras et al. [37], who mention in the paper that their model is close to our own previously proposed multi-view mammogram deep learning classifier [14,15], but their approach has been trained with 103,000 high-resolution images and results show the importance of using large training sets and high-resolution images.

The second main trend lies in the development of deep learning models that use the small training sets already available in the field. The major challenge behind this second trend is that such small datasets are rarely enough for training the high capacity deep learning models. Even though it is possible in some tasks to extract large training sets from these small datasets [20,65], for most situations, new solutions are necessary to address this issue. This is one of the most studied topics in deep learning for medical image analysis [1,4,5,14]–[16,21,33,54], where systematic studies have

[1]We thank one of the anonymous reviewers for proposing this variation.

been published [4,5,72,74,80]. The current evidence shows that transfer learning appears to address the small dataset challenge, where models can be pre-trained either in an unsupervised manner with medical image datasets or in a supervised way using non-medical image datasets. The third trend lies in the analysis of multiple input views to produce a single output - this idea has been explored in medical image analysis problems [11,14,15,17,27,37,41] and computer vision tasks [73]. Finally, the fourth trend is the holistic analysis of medical images [14,15,35,36,47,61,62,83] as opposed to the localised processing, which in general depend on the localisation and possible segmentation of structures before a classification can be achieved [1,2,22]–[26,29,31,38,46,48,63].

In this paper, the main novelty consists of the use of unregistered multi-view inputs, where images and segmentation maps are processed in a holistic manner (our original paper on the holistic analysis was also developed in parallel to the approaches cited above - note that even though we use detection of lesions, we process the whole image and not each lesion independently). We also explore the transfer learning approaches mentioned above to deal with the limited amount of training samples.

The more classic methods designed for the analysis of mammograms [38] are either based on holistic approaches that rely on traditional texture analysis [56], or on the localised analysis of lesions. The latter approach depends on a process that can be sub-divided into three stages [19,68]: 1) lesion detection, 2) lesion segmentation and 3) lesion classification. Usually, methods based on the localised analysis of lesions are limited to processing single views, but there are exceptions that work with multiple views [69]. Moreover, there have also been important developments in the exploration of deep learning models within such classical framework. In particular, the problem of detecting lesions with deep learning models has been studied with the use of large annotated training sets [48] or the use of a cascade of models and small training sets [22]. Lesion segmentation with deep learning has been addressed with the use of large training sets [29,31] or with the use of probabilistic graphical models and small training sets [23]–[25]. Finally, lesion classification methods that rely on deep learning models are generally based on a direct classification of the detected and segmented lesions [2,26,46,63]. It is important to notice that mammogram analysis systems have two goals in general: 1) the classification of an exam into normal (i.e., no findings), benign findings or malignant findings; and 2) the localisation of such findings. The sub-division adopted by classic methods is reasonable in the sense that it tries to mimic how expert radiologists work, but mathematically this sub-division makes restrictive assumptions about the problem, such as that once the analysis is focused in the lesions, the global information contained in the whole image is assumed to be irrelevant. It is also assumed that both the appearance and shape of the lesion are important for the mammogram classification process. Finally, the objective functions used for each stage form goals that are not necessarily linked with better classification – for example, the minimisation of the overlap between the annotated and detected bounding boxes is assumed to be important for classification, but never properly tested (similarly for objective functions used for segmentation).

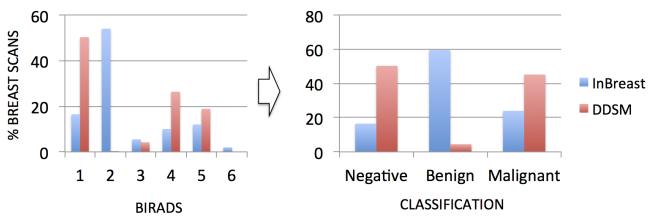


Fig. 2. Distribution of BI-RADS (left) and negative, benign and malignant classes (right) for the cases in INbreast (blue) and DDSM (red), where a case is represented by the MLO and CC mammographic views with respective segmentation maps (MCs and masses) of a single breast scan of a patient.

TABLE I
PUBLICLY AVAILABLE DATASETS USED IN THIS WORK.

Datasets:	INbreast [57]	DDSM [42]	Imagenet [66]
# Images	410	680	1.35×10^6
# Patients	115	172	-
# Classes	6	6	1000
Image Type	Mammo	Mammo	Non-medical
Annotations	BI-RADS + Lesion delineation	BI-RADS + Lesion delineation	Imagenet classes

III. Materials and Methods

In this section, we first describe the datasets and deep learning model used. Then we explain the methodological details of our approach and the experimental setup.

A. Materials

The material used in this work are the images and annotations present in the following publicly available datasets: INbreast [57], DDSM [42] and Imagenet [66] - Table I shows the number of images, patients and classes, the image type and the annotations present in each dataset. For INbreast and DDSM, a case represents the multi-view mammograms and respective segmentation maps of masses and MCs extracted from a single breast of a patient. In these two datasets, cases are manually classified into six possible Breast Imaging Reporting and Data System (BI-RADS) classes: 1) negative, 2) benign finding(s), 3) probably benign, 4) suspicious abnormality, 5) highly suggestive of malignancy, and 6) proven malignancy (see Fig. 2 for the distribution of classes in the datasets considered by this paper). It is important to note that the manual lesion delineations provided for DDSM are significantly less precise than the annotations for INbreast, as shown in Fig. 8 that presents examples of the mass and MC manual annotations from DDSM.

The INbreast and DDSM datasets are represented by $\mathcal{D} = \{(\mathbf{x}^{(p,b)}, \mathbf{c}^{(p,b)}, \mathbf{m}^{(p,b)}, \mathbf{y}^{(p,b)})\}_{p\in\{1,\dots,P\},b\in\{\text{left,right}\}}$, where $\mathbf{x} = \{\mathbf{x}_{\text{CC}}, \mathbf{x}_{\text{MLO}}\}$ denotes the CC and MLO mammography views, with $\mathbf{x}_{\text{CC}}, \mathbf{x}_{\text{MLO}} : \Omega \rightarrow \mathbb{R}$ (Ω represents the image lattice), $\mathbf{c} = \{\mathbf{c}_{\text{CC}}, \mathbf{c}_{\text{MLO}}\}$ and $\mathbf{m} = \{\mathbf{m}_{\text{CC}}, \mathbf{m}_{\text{MLO}}\}$ denote the MC and mass segmentation maps per view, with $\mathbf{c}_{\text{CC}}, \mathbf{c}_{\text{MLO}}, \mathbf{m}_{\text{CC}}, \mathbf{m}_{\text{MLO}} : \Omega \rightarrow \{0,1\}$, $\mathbf{y} \in \mathcal{Y} = \{0,1\}^C$ represents the BI-RADS classification with C classes, $p \in$

$\{1, ..., P\}$ indexes the patients, and $b \in \{\text{left}, \text{right}\}$ indexes the patient's left and right breasts (each breast is denoted as an individual case because the left and right breasts may have different BI-RADS scores). Note in Fig. 2 that these datasets have a limited amount of cases belonging to each of the six possible BI-RADS classes, so we propose a new set of three classes: 1) negative, represented by $\mathbf{y} = [1, 0, 0]^\top$, when BI-RADS=1; 2) benign, denoted by $\mathbf{y} = [0, 1, 0]^\top$, with BI-RADS $\in \{2, 3\}$; and 3) malignant, represented by $\mathbf{y} = [0, 0, 1]^\top$, when BI-RADS $\in \{4, 5, 6\}$ (the rightmost graph in Fig. 2 shows the distribution for the 3-class problems in the datasets considered by this paper). The Imagenet dataset containing non-medical images is denoted 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 $\widetilde{\mathcal{D}}$. This dataset is used for pre-training the deep learning model, as explained below in more details.

B. Deep Learning Model

The deep learning model explored in this work is the convolutional neural network (ConvNet) [50,53] CNN-F proposed by Chatfield et al. [18], which is a simplified version of the AlexNet model [50]. This model is formally denoted by $f : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ represents the image space and $\mathcal{Y}$ denotes the 3-class classification space. A ConvNet is a model containing L convolutional layers and K fully connected layers defined as follows (see Fig. 3):

$$f(\mathbf{x}; \theta) = f_{\text{out}}(f_{fc,K}(...f_{fc,1}(f_L(...f_1(\mathbf{x}, \theta_1)..., \theta_L), \theta_{fc,1})..., \theta_{fc,K}), \theta_{out}), \quad (1)$$

where $\{f_i(.)\}_{i=1}^L$ represents a convolutional layer, θ_l denotes the parameters of layer l of the ConvNet comprising the weight matrix $\mathbf{W}_l \in \mathbb{R}^{k_l \times k_l \times n_l \times n_{l-1}}$ and bias vector $\mathbf{b}_l \in \mathbb{R}^{n_l}$, with $k_l \times k_l$ denoting the size of the filters in layer l that has n_{l-1} input channels and n_l output channels (the output of this convolutional layer usually passes through a non-linear activation function and a sub-sampling stage), $f_{fc,k}$ is a fully-connected layer with weights $\{\mathbf{W}_{fc,k}\}_{k=1}^K$ (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 $\mathbf{b} \in \mathbb{R}^{n_{fc,k}}$), and f_{out} represents a multinomial logistic regression layer [50] containing the weights $\mathbf{W}_{out} \in \mathbb{R}^{n_{fc,K} \times C}$ and bias $\mathbf{b}_{out} \in \mathbb{R}^C$.

The operation in each convolutional layer $l \in \{1, ..., L\}$ of the ConvNet is defined by:

$$\mathbf{F}_l = f_l(\mathbf{x}_{l-1}, \theta_l) = \mathbf{W}_l \star \mathbf{F}_{l-1} + \mathbf{b}_l, \quad (2)$$

where $\star$ denotes the convolution operator, $\mathbf{F}_l = [\mathbf{f}_{l,1}, ..., \mathbf{f}_{l,n_l}]$, with $\mathbf{F}_0$ representing the input mammogram $\mathbf{x}$ or segmentation maps $\mathbf{c}$ or $\mathbf{m}$. Following the L^{th} convolutional layer, we have fully connected layers that take as input the vectorised input volume $\mathbf{f}_L \in \mathbb{R}^{|\mathbf{f}_L|}$ (from $\mathbf{F}_L$) (where $|\mathbf{f}_L|$ represents the length of the vector $\mathbf{f}_L$) and apply K linear transforms, defined by [50]:

$$\mathbf{f}_{fc} = f_{fc}(\mathbf{F}_L, \theta_{fc}) = (\mathbf{W}_{fc,K} ... (\mathbf{W}_{fc,1}\mathbf{f}_L + \mathbf{b}_{fc,1}) ... + \mathbf{b}_{fc,K}), \quad (3)$$

where $\mathbf{f}_{fc} \in \mathbb{R}^{n_{fc,K}}$. The final classification layer is defined by a softmax function over a linearly transformed input [50]:

$$\mathbf{f}_{out} = f_{out}(\mathbf{f}_{fc}, \theta_{out}) = softmax(\mathbf{W}_{out}\mathbf{f}_{fc} + \mathbf{b}_{out}), \quad (4)$$

where $softmax(\mathbf{z}) = \frac{e^{\mathbf{z}}}{\sum_j e^{\mathbf{z}(j)}}$, and $\mathbf{f}_{out} \in [0, 1]^C$ denotes the output from the inference process, with C representing the number of output classes. The CNN-F model [18] explored in this work, depicted in Fig. 3, has an input of 264×264×3 pixels (i.e., the input has three channels of 264×264 pixels), where the first convolutional stage has 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 has a softmax layer containing three nodes. Note that we explore two types of model inputs (see Figures 3): 1) the 2-D input model takes the input image or segmentation map and replicate it three times to fill the three input channels; and 2) the 3-D input model takes as input the image, mass and MC segmentation maps of a single view (i.e., CC or MLO) and feed them into the three input channels.

The training process for estimating $\theta = [\theta_1, ..., \theta_L, \theta_{fc,1}, ..., \theta_{fc,K}, \theta_{out}]$ in (1) is based on the minimisation of the cross entropy loss [50] over the training set, defined as [50]:

$$\ell(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{f}_{out,i}^\top \log \mathbf{y}_i, \quad (5)$$

where N represents the number of cases available for training.

C. Transfer Learning

The *pre-training stage* of the ConvNet in (1) uses the Imagenet dataset $\widetilde{\mathcal{D}}$ in order to model $\widetilde{\mathbf{y}}^* = f(\widetilde{\mathbf{x}}; \widetilde{\theta})$, where $\widetilde{\theta} = [\widetilde{\theta}_1, ..., \widetilde{\theta}_L, \widetilde{\theta}_{fc,1}, ..., \widetilde{\theta}_{fc,K}, \widetilde{\theta}_{out}]$. This pre-training is based on the minimisation of the cross-entropy loss in (5) using the $\widetilde{C}$ classes from $\widetilde{\mathcal{D}}$. This pre-trained model is then used to initialise the model parameters as follows: $\theta_1 = \widetilde{\theta}_1$, ..., $\theta_L = \widetilde{\theta}_L$, $\theta_{fc,1} = \widetilde{\theta}_{fc,1}$, ... , $\theta_{fc,K} = \widetilde{\theta}_{fc,K}$, and θ_{out} is initialised with random values (normally distributed). The fine-tuning consists of training this model by minimising the cross-entropy loss in (5) using the C classes from $\mathcal{D}$ (see Fig. 3). The motivation behind initialising almost all parameters with the pre-trained model is based on the results published by Yosinski et al. [81] that show that the success of similar fine-tuning processes depend on the use of a large number of pre-trained layers. This **fine tuning** process produces **six 2-D models** and **two 3-D models**, where the 2-D models are represented by: 1) $f(\mathbf{x}_{\text{MLO}}; \theta_{\text{MLO,im}})$, 2) $f(\mathbf{x}_{\text{CC}}; \theta_{\text{CC,im}})$, 3) $f(\mathbf{c}_{\text{MLO}}; \theta_{\text{MLO,mc}})$, 4) $f(\mathbf{c}_{\text{CC}}; \theta_{\text{CC,mc}})$, 5) $f(\mathbf{m}_{\text{MLO}}; \theta_{\text{MLO,ma}})$ and 6) $f(\mathbf{m}_{\text{CC}}; \theta_{\text{CC,ma}})$. The 3-D models are denoted by: 1) $f([\mathbf{x}_{\text{MLO}}, \mathbf{c}_{\text{MLO}}, \mathbf{m}_{\text{MLO}}]; \theta_{\text{MLO,3D}})$, and 2) $f([\mathbf{x}_{\text{CC}}, \mathbf{c}_{\text{CC}}, \mathbf{m}_{\text{CC}}]; \theta_{\text{CC,3D}})$.

D. Multi-view Analysis

The multi-view analysis of mammograms is based on merging the six 2-D and two 3-D models introduced in Sec. III-C, where we propose an evaluation that shows the performance of the classifier as a function of which layer is used to merge

the models. The process of merging the models involves the concatenation of the outputs from a particular layer, as shown in Fig. 3. In particular, we test four types of merging: 1) "JOIN 1": merge the representations $\mathbf{F}_1$ in (2) from the fine-tuned models from Sec. III-C; 2) "JOIN 2": merge the representations $\mathbf{F}_2$ in (2) from the fine-tuned models; 3) "JOIN 3": merge the representations $\mathbf{F}_L$ in (2) from the fine-tuned models; and 4) "JOIN 4": merge the representations $\mathbf{f}_{fc}$ in (3) from the fine-tuned models. After merging, the multi-view model is fine-tuned using the minimisation of the cross-entropy loss in (5) with the C classes in $\mathcal{D}$. In addition, we train four multi-view 2-D models (and another four multi-view 3-D models) using the manually defined segmentation maps and another four multi-view 2-D models (+ four multi-view 3-D models) for the automatically defined segmentation models. Given that the use of a pre-trained model can be regarded as a training regularisation approach, we compare it to another common regularisation method: data augmentation [50], obtained by randomly cropping the original training images (and respective segmentation maps) with a bounding box, whose top-left and bottom-right corners are uniformly sampled from a range of $[1, 10]$ pixels from the original corners. Note that when augmenting the data, the same transformation is applied to the mammogram view and respective mass and MC segmentations maps, but the transformations applied to the two views of the same breast may not be the same given that we do not have the registration between these two views. Therefore, all models specified above are trained with data augmentation by adding 5, 10, 20 and 50 new samples per training image.

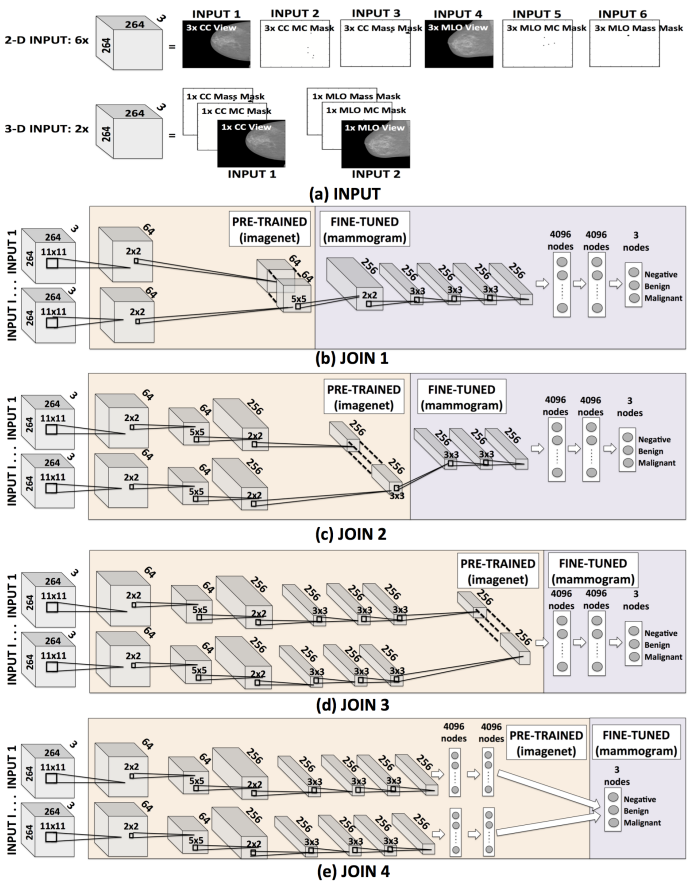


Fig. 3. Multi-view ConvNet models using different types of merging strategies for 2-D and 3-D models. The baseline model contains $L = 5$ convolutional layers, $K = 2$ fully connected layers and one final softmax layer. The model can have 2-D or 3-D inputs (a), and be pre-trained and fine-tuned in four different manners, as depicted in (b)-(e) (i.e., JOIN 1 to JOIN 4), where for the 2-D model, there are six inputs and for the 3-D model there are two inputs.

E. Automated Lesion Detection

For the automated lesion detection methods, we use recently proposed methods that produce state-of-the-art results in the INbreast dataset [57]. In particular, we use the mass detection methodology proposed by Dhungel et al. [22], consisting of a cascade of deep learning detectors that select a relatively large set of mass regions of interest (ROI), which are then processed by a cascade of random forest classifiers [10] that use appearance and shape features [78] extracted from those ROIs. For the MC detection, we use the methodology proposed by Lu et al. [55], which is based on a cascade of boosting classifiers [70] that selects ROIs containing individual MCs, where these classifiers also use appearance and shape features [78] from those ROIs. We refer the reader to those papers for more details.

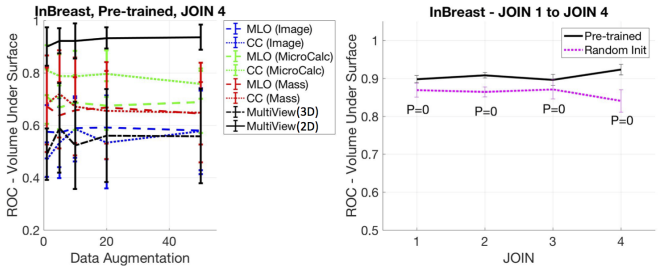
F. Experimental Setup

The input CC and MLO mammograms are pre-processed with local contrast normalisation in order to enhance the visualisation of image features and Otsu's segmentation [60] that selects a tight bounding box from the mammogram containing the breast region (see Fig. 7 for examples of the appearance of the pre-processed mammograms). The bounding box extracted from the mammogram is subsequently resized (via bi-cubic interpolation) to 264×264 pixels. We also align the input mammogram such that the pectoral muscle is always located on the right-hand side of the image. The MC and mass segmentation maps are represented by binary images that are cropped, resized and flipped in the same way as their respective mammograms.

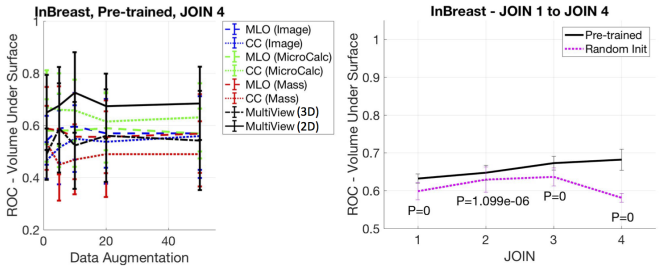
For the transfer learning experiments, all models are pre-trained [18] using the Imagenet dataset [66] that contains 1000 visual classes, 1.2×10^6 training, 50×10^3 validation and 100×10^3 test images. If transfer learning is not used, then the model parameters θ in (1) are initialised with an unbiased Gaussian with standard deviation 0.01. In all training processes, the learning rate is fixed at 0.001, momentum is equal to 0.9, weight decay is set to 0.0005, the mini-batch size is 10 and the number of epochs is 20.

The automated lesion detection experiment is run only on the INbreast dataset [57] because the manual segmentation annotations are accurate enough to allow us to build effective mass and MC detection approaches [22,55]. The imprecise manual lesion segmentation present in the DDSM dataset [42] (clearly seen in Fig. 8) does not allow the implementation of lesion detection systems [22,55], which have a detection accuracy that is high enough to allow our proposed methodology to work reliably.

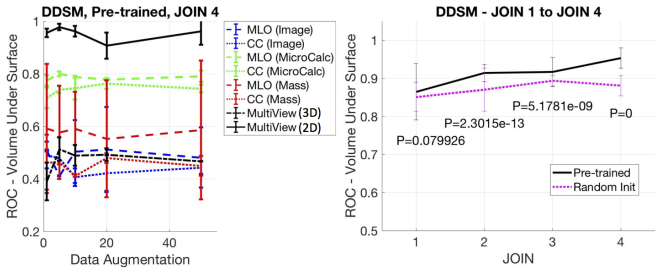
The classification accuracy is measured as follows. For a



(a) Semi-automated - INbreast

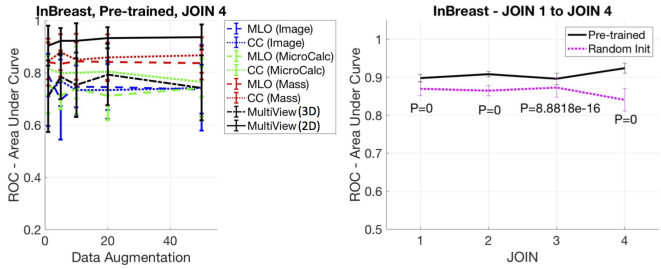


(b) Fully automated - INbreast

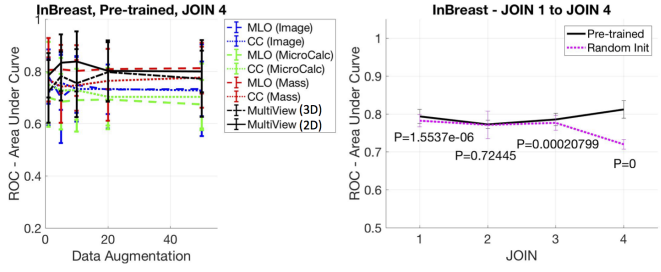


(c) Semi-automated - DDSM

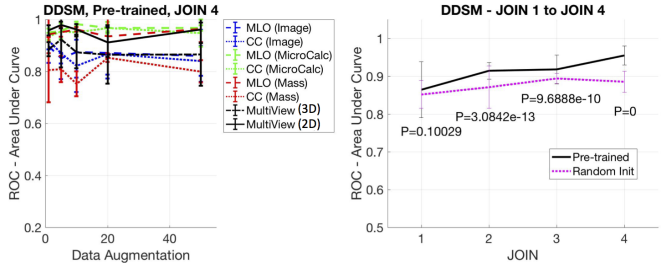
Fig. 4. **3-Class Problem:** VUS results on **INbreast** [57] (a-b) and **DDSM** [42] (c) for "JOIN 4" (pre-trained - first column) for the multi-modal (MultiView (2D) and (3D)) and individual inputs (mammographic views and segmentation maps) as a function of training data augmentation; and for all types of merging strategies (as displayed in Fig. 3) of the pre-trained and randomly initialised multi-view models using the 2-D input ("JOIN 1" to "JOIN 4" - second column). Also notice that we show the results for the semi-automated (rows a,c) and fully-automated methods (row b). The p-values show the t-test results comparing the pre-trained and randomly initialised models regarding the merging strategies and data augmentation.



(a) Semi-automated - INbreast



(b) Fully-automated - INbreast



(c) Semi-automated - DDSM

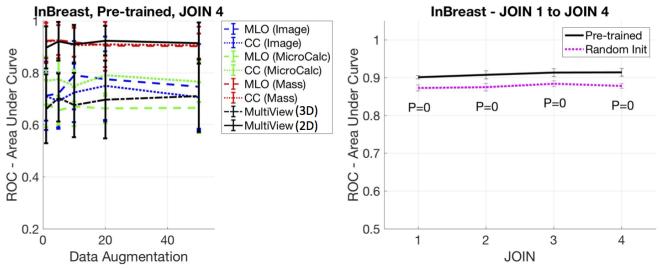
Fig. 5. **2-Class Problem - lesion classification (benign vs malignant):** AUC (lesion classification) results on **INbreast** [57] (a-b) and **DDSM** [42] (c) for "JOIN 4" (pre-trained - first column) for the multi-modal (MultiView (2D) and (3D)) and individual inputs (mammographic views and segmentation maps) as a function of training data augmentation; and for all types of merging strategies (as displayed in Fig. 3) of the pre-trained and randomly initialised multi-view models using the 2-D input ("JOIN 1" to "JOIN 4" - second column). Also notice that we show the results for the semi-automated (rows a,c) and fully-automated methods (rows b). The p-values show the t-test results comparing the pre-trained and randomly initialised models regarding the merging strategies and data augmentation.

3-class problem, with classes negative, benign and malignant, the accuracy is measured with the volume under ROC surface (VUS) [51]. The lesion classification "benign vs malignant" 2-class problem is assessed with the area under ROC curve (AUC), where it is assumed that all cases contain at least one finding (i.e., an MC or a mass). The breast screening "malignant vs benign/normal" 2-class problem is also assessed with AUC. We assess the semi-automated method (using the manually defined segmentation maps of masses and MCs) on both datasets, and the fully-automated method (using the automatically defined segmentation maps of lesions explained in Sec. III-E) on INbreast. Finally, for the INbreast dataset, results are computed from a 5-fold cross validation experiment, where each fold consists of a training set containing 90 patients and a testing set with 25 patients. For DDSM, the results are calculated using the suggested division of training and testing sets for DDSM [42], with 86 patients for training and 86 for testing - this allows a direct comparison with previously reported results. All statistical significance tests are based on the unpaired t-test.

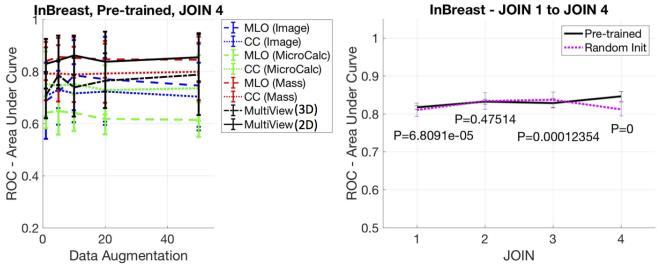
IV. Results

Before presenting the results of our proposed methodology, we summarised the results of the lesion detection systems presented in Sec. III-F. For the INbreast dataset, using 5-fold cross validation experiment, the automated MC detection [55] can detect 40% of the MCs at one false positive per image (FPI), and 80% of the MCs at 10 FPI, while the mass detection [22] can detect around 96% of the masses at around 1 FPI and more than 98% of the masses at 10 FPI. The operating point for both detectors was chosen in order to have on average 1 FPI.

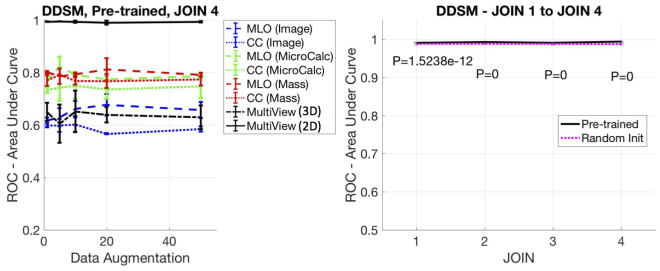
The assessment of our approach is depicted in Figures 4-6. We focus the explanation of the results with respect to the following points: 1) individual versus multi-modal inputs, 2)



(a) Semi-automated - INbreast



(b) Fully-automated - INbreast



(c) Semi-automated - DDSM

Fig. 6. **2-Class problem- breast screening (negative/normal findings vs malignant):** AUC (lesion classification) results on **INbreast** [57] (a-b) and **DDSM** [42] (c) for "JOIN 4" (pre-trained - first column) for the multi-modal (MultiView (2D) and (3D)) and individual inputs (mammographic views and segmentation maps) as a function of training data augmentation; and for all types of merging strategies (as displayed in Fig. 3) of the pre-trained and randomly initialised multi-view models using the 2-D input ("JOIN 1" to "JOIN 4" - second column). Also notice that we show the results for the semi-automated (rows a,c) and fully-automated methods (rows b). The p-values show the t-test results comparing the pre-trained and randomly initialised models regarding the merging strategies and data augmentation.

different types of merging strategies (JOIN 1 to 4), 3) pre-trained versus randomly initialised models, and 4) fully- versus semi-automated methods. We also show several visual results in Figures 7 and 8.

The first column of Figures 4-6 show the VUS (3-class problem) and AUC (2-class problems) results on INbreast and DDSM for the **individual inputs (CC and MLO views, mass and MC segmentation maps) and the multi-modal input (labelled as multi-view)**. For the majority of the cases, the main evidence noticed is that the multi-modal input produces classification results that tend to be at least as good as the best result from the individual inputs. Furthermore, the second columns of Figures 4-6 show the results produced by the **different types of merging strategies (JOIN 1 to 4)** of the Multi-view ConvNet models averaged over the all different amounts of data augmentation considered in this work (1, 5, 10, 20, and 50). In all cases, it is clear that the JOIN 4 strategy for the pre-trained model produces the best overall results. For the randomly initialised model, the trend is slightly different with the JOIN 1,2,3 strategies producing similar results and JOIN 4 with a slightly worse performance. It is also noticeable that the **pre-trained model is consistently better than the randomly initialised counterpart**, as evidenced by the statistically significant t-test results. Finally, another important point to notice is the **difference in performance between the fully and semi-automated method**, shown in Figures 4-6 for the INbreast dataset. More specifically, row (a) of these figures display the semi-automated methods and rows (b) show the fully automated cases (on INbreast). There is a significant performance deterioration for the 3-class problem, which is less significant for the 2-class problems.

In Tab. IV, we compare our results to the latest state-of-the-art (SoA) results published in the field [27,37,83,83] - note that all these SoA results have been published after our original publication [15], but before the submission of our revised manuscript, so these SoA methods have been developed either in parallel or after our own approach. The t-test between our proposed method and Dhungel et al. [27] for breast screening using automated detected lesions on INbreast shows $p < 1 \times 10^{-10}$, and $p > 0.05$ when relying on manually detected lesions. Also, the t-test between our proposed method and Zhu et al. [83] for breast screening using automated detected lesions on INbreast shows $p > 0.05$ (in fact, Zhu et al. [83]'s approach does not need an automated lesion detection stage - but they can detect lesion as a side effect of their approach). We cannot compare directly the meanAUC results between our approach and Geras et al. [37] because they have been obtained from different datasets and different classification problems (classes are different - see column "Classes"). Finally, using our proposed model, we also compute the specificity and sensitivity results for the breast screening problem using the operating point closest to the equal error rate on the ROC curves. For INbreast, using manually detected lesions, we have $sp = 0.92 \pm 0.08$ and $se = 0.69 \pm 0.28$, and using automatically detected lesions, we have $sp = 0.66 \pm 0.14$ and $se = 0.69 \pm 0.23$. For DDSM, using manually detected lesions, we have $sp = 0.97 \pm 0.01$ and $se = 0.94 \pm 0.01$.

We also show several correctly and incorrectly (Fig. 7) classified test cases from INbreast produced by the fully-automated method (JOIN 4) trained with 50× data augmentation, and test cases from DDSM (Fig. 8) produced by the semi-automated method (JOIN 4) trained with 50× data augmentation. Finally, running Matconvnet [18] on a standard desktop (2.3GHz Intel Core i7 with 8GB, and graphics card NVIDIA GeForce GT 650M 1024 MB), the time for training six models and the multi-view model (without data augmentation) is one hour. With the addition of 10× data augmentation, the training time increases to four hours, with 20 × data augmentation, the training time increases to seven hours, and with 50 × data augmentation, the training time increases to over 12 hours.

V. Discussion

Figures 4-6 show that our proposed approach can jointly classify unregistered and multi-modal (images and segmentation maps) inputs using high-level deep learning features. In

TABLE II
THIS TABLE DISPLAYS FOR EACH SOA METHOD (PROPOSED: PRE-TRAINED, JOIN4, WITH $50\times$ DATA AUGMENTATION), THE DATASET (COLUMNS 2-5), IF IT IS FULLY AUTOMATED (COLUMN "AUTO"), THE 3-CLASS (VUS) RESULT, THE LESION CLASSIFICATION (LC) AND BREAST SCREENING (BS) 2-CLASS RESULTS, AND THE MEANAUC RESULTS (THE MEANAUC [37] IS COMPUTED BY TAKING THE AVERAGE OF THREE CLASSIFICATION PROBLEMS BASED ON MAKING ONE OF THE THREE CLASSES POSITIVE AND THE OTHER TWO, NEGATIVE). THE SYMBOL '?' INDICATES THAT THE RESULT IS NOT PUBLICLY AVAILABLE.

Method	Dataset	# cases	# images	Classes	Auto	VUS	AUC (LC)	AUC (BS)	meanAUC
Proposed	INbreast	115	410	$\{Neg., Ben., Mal.\}$	NO	0.94 ± 0.05	0.94 ± 0.05	0.91 ± 0.08	0.87 ± 0.08
Proposed	DDSM	172	680	$\{Neg., Ben., Mal.\}$	NO	0.96 ± 0.05	0.96 ± 0.05	0.99 ± 0.01	0.91 ± 0.03
Dhungel et al. [27]	INbreast	115	410	$\{Neg., Ben., Mal.\}$	NO	?	?	0.91 ± 0.03	?
Proposed	INbreast	115	410	$\{Neg., Ben., Mal.\}$	YES	0.68 ± 0.14	0.78 ± 0.09	0.86 ± 0.09	0.72 ± 0.10
Dhungel et al. [27]	INbreast	115	410	$\{Neg., Ben., Mal.\}$	YES	?	?	0.80 ± 0.04	?
Zhu et al. [83]	INbreast	115	410	$\{Neg., Ben., Mal.\}$	YES	?	?	0.86 ± 0.03	?
Geras et al. [37]	Private	$\approx 18K$	$\approx 100K$	$BIRADS \in \{0, 1, 2\}$	YES	?	?	?	0.69

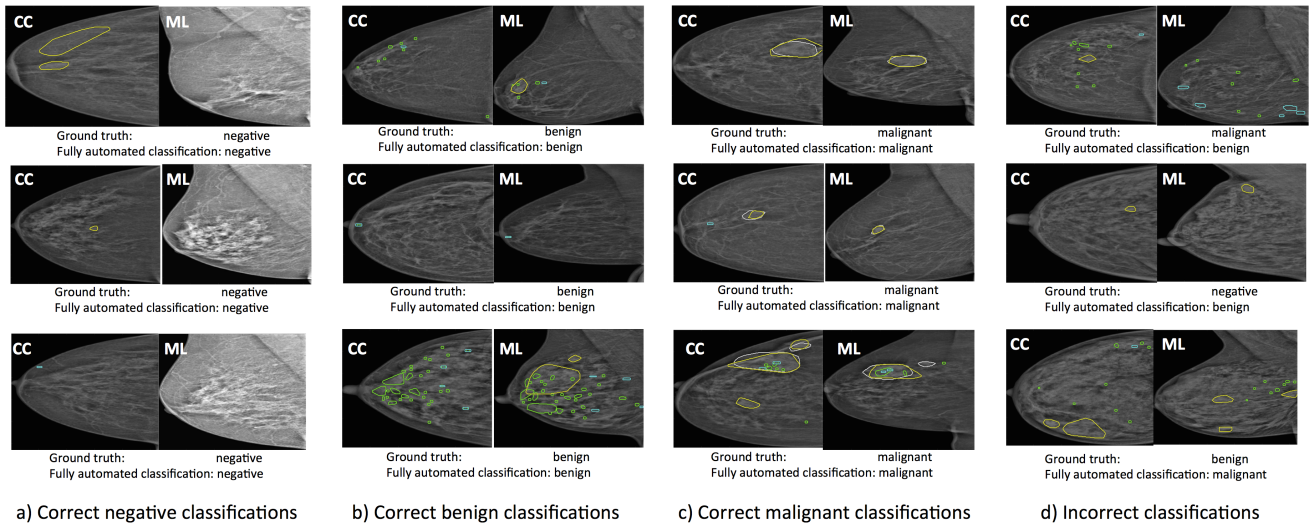


Fig. 7. Correct (a-c) and incorrect (d) classifications on INbreast [57] test cases using Imagenet pre-trained "JOIN 4" model with $50\times$ data augmentation, where the ground truth (GT) and the automatic (AUTO) mass (GT in white, AUTO in yellow) and MC (GT in green, AUTO in cyan) detections and classifications (text below images) are shown.

particular, it is important to observe that amongst pre-trained models, "JOIN 4" presents better results than "JOIN 1,2,3", suggesting that higher level features in the deep learning model are more appropriate to achieve the goal of multi-view classification. Similar conclusions have been achieved by previous approaches that showed that the merging of deep learning models is more effective when they are joined at the high-level layers [8]. Moreover, for the randomly initialised models, "JOIN 4" performs worse than "JOIN 1,2,3", suggesting that the larger number of parameters present in that model (as shown in Fig. 3) makes the use of pre-trained models more critical. Another important point shown in Figures 4-6 is that all results indicate that the use of the 3-D input does not lead to competitive classification results - this may happen because these networks are likely to need channels containing highly correlated data, but this is a topic that needs more study in future works. Furthermore, Figures 4-6 suggest that pre-trained models lead to statistically significant improvements compared to the randomly initialised ones.

The little difference between the VUS for the 3-class problem and AUC for the 2-class problem (benign vs malignant) in the semi-automated model can be explained by the fact that the proposed model is nearly perfect in classifying cases that do not contain any findings, demonstrating the ability of the model to classify an input without lesions as negative - a high level classification challenge. Furthermore, given the false positive detections produced by the automated lesion detectors, the fully-automated model must try to classify negative cases even with the presence of false positive mass or MC detections, which is shown to happen in Fig. 7. From Figures 4-6, we see that the results for the fully-automated multi-view pre-trained models is better than the individual results for the 3-class problem, but on par with the best individual input for the 2-class problems. This also indicates robustness to the false positive detections of the automated lesion detectors, but it also shows that such false positive detections have a negative impact in the ability of the model to classify correctly a whole exam in a holistic manner. In general, the models show poor performance for the single view classifications, which may happen because cases where BI-RADS > 1 may contain annotations for either MC or mass, but not for both lesions. Moreover, mammographic views (CC and MLO) may have

insufficient information for a robust classification, especially considering that they are down-sampled around ten times from their original size. Finally, in the 3-class problem, the MC segmentation maps produce better classification results in isolation than mass maps, which in turn are better than the mammograms, while for 2-class problems, mass maps tend to produce better classification results than MC maps. This is evidence that these segmentation maps have different roles depending on the classification problem being studied.

The comparison with SoA methods in Tab. IV shows that our approach is competitive with methods effectively published after our own original publication [15]. We can also compare our approach with previously published semi-automated lesion classification methods [38], which produce and AUC in $[0.9, 0.95]$ for MCs and mass classification [19, 79]. Hence, our proposed method is competitive on INbreast (our AUC in $[0.9, 0.98]$) and superior on DDSM (our AUC in $[0.91, 1.0]$) with respect to these approaches. In addition, we can also compare our results to more recently proposed methods based on deep learning. Dhungel et al.'s semi-automated mass classification method [26] has produced an AUC = 0.91, and the fully automated has yielded an AUC = 0.76 on INbreast - these results are comparable, but slightly worse than our results. Finally, the sensitivity and specificity results shown by our proposed model is competitive to the results produced by radiologists in breast screening classification when we rely on manual lesion annotation. However, when considering the fully automated method, ours and current SoA methods still need to match human performance.

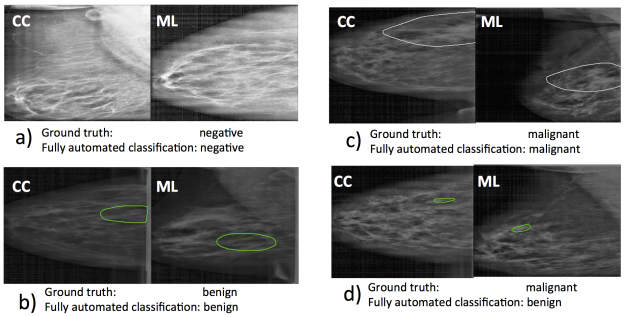


Fig. 8. Correct (a-c) and incorrect (d) classifications on DDSM [42] test cases using Imagenet pre-trained "JOIN 4" model with $50\times$ data augmentation, where the ground truth mass (white) and MC (green) detections and classifications (text below images) are shown. Notice that the manual mass and MC annotations are significantly less precise than the ones from INbreast, shown in Fig. 7.

VI. Conclusion and Future Work

In this paper, we demonstrate that high-level deep learning features can be used in the classification of unregistered multi-view and multi-modal input mammograms and segmentation maps. The use of such deep learning models is facilitated by the use of models that have been pre-trained with computer vision datasets containing millions of non-medical images. Our results shown in Sec. IV can be used as a benchmark in the field given that both datasets are publicly available. We believe that our proposed work introduces an important research topic to the field: the analysis of un-registered multi-view and multi-modal medical images. We plan to extend our proposed approach in the following directions: 1) make it robust to the large number of false positives produced by the automated lesion detections, 2) remove the dependence on manual lesion annotations for training the deep learning model and rely only on the annotations available from the clinical dataset [83] (e.g., mammogram classification, radiology reports, and patient data), 3) use large scale datasets containing high-resolution images [37], and 4) combine different breast imaging modalities.

References

[1] A. Akselrod-Ballin, L. Karlinsky, S. Alpert, S. Hasoul, R. Ben-Ari, and E. Barkan. A region based convolutional network for tumor detection and classification in breast mammography. In *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 197–205. Springer, 2016.

[2] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 2016.

[3] J. E. Ball and L. M. 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*, pages 4973–4978. IEEE, 2007.

[4] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan. Chest pathology detection using deep learning with non-medical training. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 294–297. IEEE, 2015.

[5] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan. Chest pathology identification using deep feature selection with non-medical training. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–5, 2016.

[6] M. Beller, R. Stotzka, T. O. Müller, and H. Gemmeke. An example-based system to support the segmentation of stellate lesions. In *Bildverarbeitung für die Medizin 2005*, pages 475–479. Springer, 2005.

[7] R. Bellotti, F. De Carlo, S. Tangaro, G. Gargano, G. Maggipinto, M. Castellano, R. Massafra, D. Cascio, F. Fauci, R. Magro, et al. A completely automated cad system for mass detection in a large mammographic database. *Medical physics*, 33(8):3066–3075, 2006.

[8] Y. Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[9] C. M. Bishop. Pattern recognition. *Machine Learning*, 2006.

[10] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[11] T. Brosch, Y. Yoo, D. K. Li, A. Traboulsee, and R. 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*, pages 462–469. Springer, 2014.

[12] R. Campanini, D. Dongiovanni, E. Iampieri, N. Lanconelli, M. Masotti, G. Palermo, A. Riccardi, and M. Roffilli. A novel featureless approach to mass detection in digital mammograms based on support vector machines. *Physics in Medicine and Biology*, 49(6):961, 2004.

[13] J. S. Cardoso, I. Domingues, and H. 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):1555002, 2015.

[14] G. Carneiro, J. Nascimento, and A. Bradley. Deep learning models for classifying mammogram exams containing unregistered multi-view images and segmentation maps of lesions. book chapter. In *in Deep Learning for Medical Image Analysis*. Elsevier, 2017.

[15] G. Carneiro, J. Nascimento, and A. P. Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In *Medical Image Computing and Computer-Assisted InterventionMICCAI 2015*, pages 652–660. Springer, 2015.

[16] G. Carneiro and J. 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):2592–2607, 2013.

[17] G. Carneiro, T. Peng, C. Bayer, and N. Navab. Automatic quantification of tumour hypoxia from multi-modal microscopy images using weakly-supervised learning methods. *IEEE Transactions on Medical Imaging*, 2017.

[18] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[19] H. Cheng, X. Shi, R. Min, L. Hu, X. Cai, and H. Du. Approaches for automated detection and classification of masses in mammograms. *Pattern recognition*, 39(4):646–668, 2006.

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

[21] A. A. Cruz-Roa, J. E. A. Ovalle, A. Madabhushi, and F. A. G. 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*, pages 403–410. Springer, 2013.

[22] N. Dhungel, G. Carneiro, and A. 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*, pages 1–8, Nov 2015.

[23] N. Dhungel, G. Carneiro, and A. P. Bradley. Deep learning and structured prediction for the segmentation of mass in mammograms. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pages 605–612. Springer, 2015.

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

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

[26] N. Dhungel, G. Carneiro, and A. P. Bradley. The automated learning of deep features for breast mass classification from mammograms. In *Medical Image Computing and Computer-Assisted InterventionMICCAI 2016*. Springer, 2016.

[27] N. Dhungel, G. Carneiro, and A. P. Bradley. Fully automated classification of mammograms using deep residual neural networks. In *2017 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, April 2017.

[28] I. Domingues, E. Sales, J. Cardoso, and W. Pereira. Inbreast-database masses characterization. *XXIII CBEB*, 2012.

[29] A. Dubrovina, P. Kisilev, B. Ginsburg, S. Hashoul, and R. Kimmel. Computational mammography using deep neural networks. In *Workshop on Deep Learning in Medical Image Analysis (DLMIA)*, 2016.

[30] N. H. Eltonsy, G. D. Tourassi, and A. S. Elmaghraby. A concentric morphology model for the detection of masses in mammography. *Medical Imaging, IEEE Transactions on*, 26(6):880–889, 2007.

[31] M. G. Ertosun and D. L. Rubin. Probabilistic visual search for masses within mammography images using deep learning. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 1310–1315. IEEE, 2015.

[32] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[33] R. Fakoor, F. Ladhak, A. Nazi, and M. 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.

[34] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1915–1929, 2013.

[35] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers, et al. Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–6, 2016.

[36] M. Gao, Z. Xu, L. Lu, A. P. Harrison, R. M. Summers, and D. J. Mollura. Multi-label deep regression and unordered pooling for holistic interstitial lung disease pattern detection. In *International Workshop on Machine Learning in Medical Imaging*, pages 147–155. Springer, 2016.

[37] K. J. Geras, S. Wolfson, S. Kim, L. Moy, and K. Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047*, 2017.

[38] M. L. Giger, N. Karssemeijer, and J. A. Schnabel. Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. *Annual review of biomedical engineering*, 15:327–357, 2013.

[39] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014.

[40] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

[41] Y. Guo, G. Wu, L. A. Commander, S. Szary, V. Jewells, W. Lin, and D. Shen. Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, pages 308–315. Springer, 2014.

[42] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer. The digital database for screening mammography. In *Proceedings of the 5th international workshop on digital mammography*, pages 212–218, 2000.

[43] K. Horsch, M. L. Giger, C. J. Vyborny, L. Lan, E. B. Mendelson, and R. E. Hendrick. Classification of breast lesions with multimodality computer-aided diagnosis: Observer study results on an independent clinical data set. *Radiology*, 240(2):357–368, 2006.

[44] M. M. Jadoon, Q. Zhang, I. U. Haq, S. Butt, and A. Jadoon. Three-class mammogram classification based on descriptive cnn features. *BioMed Research International*, 2017, 2017.

[45] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun. Cancer statistics, 2008. *CA: a cancer journal for clinicians*, 58(2):71–96, 2008.

[46] Z. Jiao, X. Gao, Y. Wang, and J. Li. A deep feature based framework for breast masses classification. *Neurocomputing*, 2016.

[47] M. Kallenberg, K. Petersen, M. Nielsen, A. Ng, P. Diao, C. Igel, C. Vachon, K. Holland, N. Karssemeijer, and M. Lillholm. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. 2016.

[48] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, 35:303–312, 2017.

[49] E. Kozegar, M. Soryani, B. Minaei, I. Domingues, et al. Assessment of a novel mass detection algorithm in mammograms. *Journal of cancer research and therapeutics*, 9(4):592, 2013.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, volume 1, page 4, 2012.

[51] T. C. Landgrebe and R. P. Duin. Efficient multiclass roc approximation by decomposition via confusion matrix perturbation analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(5):810–822, 2008.

[52] B. Lauby-Secretan, C. Scoccianti, D. Loomis, L. Benbrahim-Tallaa, V. Bouvard, F. Bianchini, and K. Straif. Breast-cancer screeningviewpoint of the iarc working group. *New England Journal of Medicine*, 372(24):2353–2358, 2015.

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

[54] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji. Deep learning based imaging data completion for improved brain disease diagnosis. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, pages 305–312. Springer, 2014.

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

[56] A. Manduca, M. J. Carston, J. J. Heine, C. G. Scott, V. S. Pankratz, K. R. Brandt, T. A. Sellers, C. M. Vachon, and J. R. Cerhan. Texture features from mammographic images and risk of breast cancer. *Cancer Epidemiology and Prevention Biomarkers*, 18(3):837–845, 2009.

[57] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic Radiology*, 19(2):236–248, 2012.

[58] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[59] A. Oliver, J. Freixenet, J. Marti, E. Perez, J. Pont, E. R. Denton, and R. Zwiggelaar. A review of automatic mass detection and segmentation in mammographic images. *Medical Image Analysis*, 14(2):87–110, 2010.

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

[61] K. Petersen, M. Nielsen, P. Diao, N. Karssemeijer, and M. Lillholm. Breast tissue segmentation and mammographic risk scoring using deep learning. In *Breast Imaging*, pages 88–94. Springer, 2014.

[62] Y. Qiu, Y. Wang, S. Yan, M. Tan, S. Cheng, H. Liu, and B. 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*, pages 978521–978521. International Society for Optics and

Photonics, 2016.
[63] Y. Qiu, S. Yan, M. Tan, S. Cheng, H. Liu, and B. Zheng. Computer-aided classification of mammographic masses using the deep learning technology: a preliminary study. In *SPIE Medical Imaging*, pages 978520–978520. International Society for Optics and Photonics, 2016.
[64] P. Rahmati, A. Adler, and G. Hamarneh. Mammography segmentation with maximum likelihood active contours. *Medical image analysis*, 16(6):1167–1186, 2012.
[65] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. 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*, pages 520–527. Springer, 2014.
[66] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014.
[67] M. P. Sampat, A. C. Bovik, G. J. Whitman, and M. K. Markey. A model-based framework for the detection of spiculated masses on mammographya). *Medical physics*, 35(5):2110–2123, 2008.
[68] M. P. Sampat, M. K. Markey, A. C. Bovik, et al. Computer-aided detection and diagnosis in mammography. *Handbook of image and video processing*, 2(1):1195–1217, 2005.
[69] M. Samulski and N. Karssemeijer. Optimizing case-based detection performance in a multiview cad system for mammography. *IEEE Transactions on Medical Imaging*, 30(4):1001–1009, 2011.
[70] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. Rusboost: improving classification performance when training data is skewed. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.
[71] J. Shi, B. Sahiner, H.-P. Chan, J. Ge, L. Hadjiiski, M. A. Helvie, A. Nees, Y.-T. Wu, J. Wei, C. Zhou, et al. Characterization of mammographic masses based on level set segmentation with new image features and patient information. *Medical physics*, 35(1):280–290, 2008.
[72] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.
[73] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.
[74] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.
[75] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa, and Y. Yang. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *Information Technology in Biomedicine, IEEE Transactions on*, 13(2):236–251, 2009.
[76] G. M. te Brake, N. Karssemeijer, and J. H. Hendriks. An automatic method to discriminate malignant masses from normal tissue in digital mammograms. *Physics in Medicine and Biology*, 45(10):2843, 2000.
[77] C. Varela, S. Timp, and N. Karssemeijer. Use of border information in the classification of mammographic masses. *Physics in Medicine and Biology*, 51(2):425, 2006.
[78] J. Wei, B. Sahiner, L. M. Hadjiiski, H.-P. Chan, N. Petrick, M. A. Helvie, M. A. Roubidoux, J. Ge, and C. Zhou. Computer-aided detection of breast masses on full field digital mammograms. *Medical physics*, 32(9):2827–2838, 2005.
[79] L. Wei, Y. Yang, R. M. Nishikawa, and Y. Jiang. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. *Medical Imaging, IEEE Transactions on*, 24(3):371–380, 2005.
[80] Z. Xu, S. Huang, Y. Zhang, and D. Tao. Webly-supervised fine-grained visual categorization via deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.
[81] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.
[82] Y. Zhang, K. Sohn, R. Villegas, G. Pan, and H. Lee. Improving object detection with deep convolutional networks via bayesian optimization and structured prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 249 – 258, june 2015.
[83] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. *arXiv preprint arXiv:1612.05968*, 2016.