Unregistered Multiview Mammogram Analysis with Pre-trained Deep Learning Models

Gustavo Carneiro[1] Jacinto Nascimento[2] Andrew P. Bradley[3]

[1] ACVT, University of Adelaide, Australia*
[2] ISR, Instituto Superior Tecnico, Portugal*
[3] University of Queensland, Australia *

Abstract. We show two important findings on the use of deep convolutional neural networks (CNN) in medical image analysis. First, we show that CNN models that are pre-trained using computer vision databases (e.g., Imagenet) are useful in medical image applications, despite the significant differences in image appearance. Second, we show that multiview classification is possible without the pre-registration of the input images. Rather, we use the high-level features produced by the CNNs trained in each view separately. Focusing on the classification of mammograms using craniocaudal (CC) and mediolateral oblique (MLO) views and their respective mass and micro-calcification segmentations of the same breast, we initially train a separate CNN model for each view and each segmentation map using an Imagenet pre-trained model. Then, using the features learned from each segmentation map and unregistered views, we train a final CNN classifier that estimates the patient's risk of developing breast cancer using the Breast Imaging-Reporting and Data System (BI-RADS) score. We test our methodology in two publicly available datasets (InBreast and DDSM), containing hundreds of cases, and show that it 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. In general, our approach shows state-of-the-art classification results and demonstrates a new comprehensive way of addressing this challenging classification problem.

Keywords: Deep learning, Mammogram, Multiview classification

1 Introduction

Deep learning models are producing quite competitive results in computer vision and machine learning [1], but the application of such large capacity models in medical image analysis is complicated by the fact that they need large training sets that are rarely available in our field. This issue is circumvented in computer vision and machine learning with the use of publicly available pre-trained deep learning models, which are estimated with large annotated databases [2] and re-trained (or fine-tuned) for other problems that contain smaller annotated training sets. This fine-tuning process has been shown to improve the generalization of the model, compared to a model that is trained with randomly initialized weights using only the small datasets [1]. However, such pre-trained models

* 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).

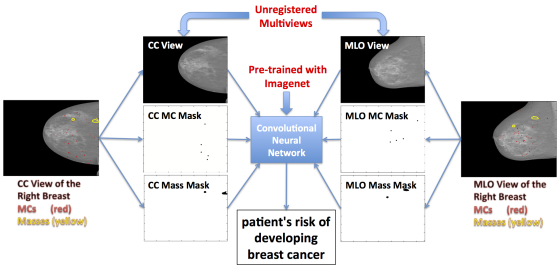


Fig. 1: Model proposed in this paper using unregistered CC/MLO views and MC/Mass segmentations of the same breast, where the classification of the patient's risk of developing breast cancer uses a CNN pre-trained with Imagenet [2].

are not available for medical image applications, so an interesting question is if models pre-trained in other (non-medical) datasets are useful in medical imaging applications. Another important potential advantage provided by deep learning methods is the automatic estimation of useful features that provide a high-level representation of the input data, which is robust to image transformations [3]. A pertinent question with regards to this point is if such features can be used in multiview classification without the need to pre-register the input views.

Literature Review: Deep learning has become one of the most exciting topics in computer vision and machine learning [4]. The main advantage brought by deep learning models, and in particular by deep convolutional neural networks (CNN) [3], is the high-level features produced by the top layers of the model that are shown to improve classification results, compared to previous results produced by hand-built features [4]. Moreover, these high-level features have also been shown to be robust to image transformations [5]. Nevertheless, the training process for CNNs requires large amounts of annotated samples (usually, in the order of 100Ks) to avoid overfitting to the training data given the large model capacity. This issue has been handled with transfer learning, which retrains (in a process called fine-tuning) publicly available models (pre-trained with large datasets) using smaller datasets [1]. However, it is still not possible to utilize this approach in medical image analysis given the lack of large publicly available training sets or pre-trained models.

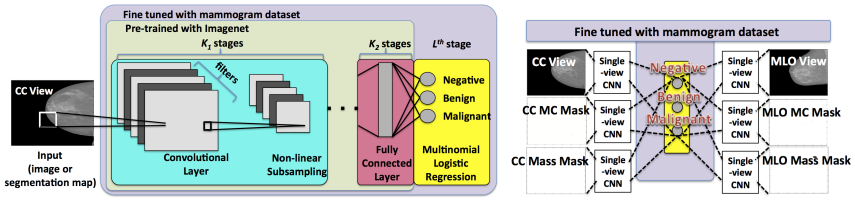
In spite of the issues above, we have seen the successful use of deep learning in medical imaging. Recently, Cireşan et al. [6] and Roth et al. [7] set up their training procedures in order to robustly estimate CNN model parameters, and in both cases (mitosis and lymph node detection, respectively), their results suparssed the current state of the art by a large margin. Another way of handling the issues above is with unsupervised training of deep autoencoders that are fine-tuned to specific classification problems [8–11]. Other interesting approaches using autoencoders for processing multiview data have been proposed [12, 13], but they require pre-registered input data. However, to date we have not seen the use of pre-trained CNN models in medical imaging, and we have not seen the analysis of unregistered input images in multiview applications with CNN models.

The literature concerning the classification of the patient's risk of developing breast cancer using mammograms is quite vast, and cannot be fully covered here, so we focus on the most relevant works for our paper. In a recent survey, Giger et al. [14] describe recent methodologies for breast image classification, but these methods focus only on binary classification of microcalcification (MC) and mass like lesions. This is different from the more comprehensive approach taken here, where we aim at classifying a whole breast exam. The state-of-the-art binary classification of masses and MCs into benign/malignant [15, 16] produces an area under the receiver operating characteristic (ROC) curve between $[0.9, 0.95]$, but these results cannot be used as baseline because they usually use databases and annotations that are not publicly available. More similar to our approach, the multimodal analysis that takes lesions imaged from several modalities (e.g., mammograms and sonograms) have been shown to improve the average performance of radiologists [17]. We believe that the new approach here proposed based on the combination of multiview analysis with MC and mass detection has the potential to improve the overall breast exam classification in terms of sensitivity and specificity.

Contributions: We show two results in this paper (see Fig. 1). First, we show that CNN models pre-trained with a typical computer vision database [2] can be used to boost classification results in medical image analysis problems. Second, we show that the high-level features produced by such CNN models allow the classification of unregistered multiview data. We propose a methodology that takes unregistered CC/MLO mammograms and MC/Mass segmentations and produce a classification based on Breast Imaging-Reporting and Data System (BI-RADS) score. Using two publicly available databases [18, 19], containing a total of 287 patients and 1090 images, we show the benefits of these two contributions and also show that we demonstrate improved classification performance for this problem. Finally, given our use of publicly available datasets only, these results can be used as baseline for future proposals.

2 Methodology

Assume we have a dataset $\mathcal{D} = \{(\mathbf{x}^{(p,b)}, \mathbf{c}^{(p,b)}, \mathbf{m}^{(p,b)}, \mathbf{y}^{(p,b)})\}_{p,b}$, where $\mathbf{x} = \{\mathbf{x}_{\text{CC}}, \mathbf{x}_{\text{MLO}}\}$ denotes the two views (CC and MLO) available (with $\mathbf{x}_{\text{CC}}, \mathbf{x}_{\text{MLO}} : \Omega \to \mathbb{R}$ and Ω denoting the image lattice), $\mathbf{c} = \{\mathbf{c}_{\text{CC}}, \mathbf{c}_{\text{MLO}}\}$ is the MC segmentation in each view with $\mathbf{c}_{\text{CC}}, \mathbf{c}_{\text{MLO}} : \Omega \to \{0, 1\}$, $\mathbf{m} = \{\mathbf{m}_{\text{CC}}, \mathbf{m}_{\text{MLO}}\}$ represents the mass segmentation in each view with $\mathbf{m}_{\text{CC}}, \mathbf{m}_{\text{MLO}} : \Omega \to \{0, 1\}$, $\mathbf{y} \in \mathcal{Y} = \{0, 1\}^C$ denotes 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 patient's breast is denoted as a case because they can be labeled different BI-RADS scores). The BI-RADS classification has 6 classes (1: negative, 2: benign finding(s), 3: probably benign, 4: suspicious abnormality, 5: highly suggestive of malignancy, 6: proven malignancy), but given the small amount of training data in some of these classes in the publicly available datasets (Fig. 3), we divide these original classes into three categories: negative or $\mathbf{y} = [1, 0, 0]^\top$ (BI-RADS=1), benign or $\mathbf{y} = [0, 1, 0]^\top$ (BI-RADS $\in \{2, 3\}$) and malignant or $\mathbf{y} = [0, 0, 1]^\top$ (BI-RADS $\in \{4, 5, 6\}$). We also have a dataset of non-medical image data for



a) Single view CNN model b) Multiview CNN model

Fig. 2: Visualization of the single view (a) and multiview (b) CNN models with K_1 stages of convolutional and non-linear sub-sampling layers, K_2 stages of fully connected layers and one final layer of the multinomial logistic regressor.

pre-training the CNN, $\widetilde{\mathcal{D}} = \{(\widetilde{\mathbf{x}}^{(n)}, \widetilde{\mathbf{y}}^{(n)})\}_n$, with $\widetilde{\mathbf{x}} : \Omega \rightarrow \mathbb{R}$ and $\widetilde{\mathbf{y}} \in \widetilde{\mathcal{Y}} = \{0,1\}^{\widetilde{C}}$ (i.e., the set of $\widetilde{C}$ classes present in dataset $\widetilde{\mathcal{D}}$).

Convolutional Neural Network: A CNN model consists of a network with multiple processing stages, each comprising two layers (the convolutional layer, where the filters are applied to the image, and the non-linear subsampling layer, which reduces the input image size), followed by several fully connected layers and a multinomial logistic regression layer [4] (Fig. 2(a)). The convolution layers compute the output at location j from input at i using the filter $\mathbf{W}_k$ and bias b_k (at k^{th} stage) using $\mathbf{x}_k(j) = \sigma(\sum_{i \in \Omega(j)} \mathbf{x}_{k-1}(i) * \mathbf{W}_k(i,j) + b_k(j))$, where $\sigma(.)$ is a non-linear function [4] (e.g., logistic or rectification linear unit), $*$ represents the convolution operator, and $\Omega(j)$ is the input region addresses; while the non-linear subsampling layers are defined by $\mathbf{x}_k(j) = \downarrow (\mathbf{x}_{k-1}(j))$, where $\downarrow (.)$ denotes a subsampling function that pools (using the mean or max functions) the values from the region $\Omega(j)$ of the input data. The fully connected layers consist of the convolution equation above using a separate filter for each output location, using the whole input from the previous layer, and the multinomial logistic regression layer computes the probability of the i^{th} class using the features $\mathbf{x}_L$ from the L^{th} layer with the softmax function $\mathbf{y}(i) = \frac{e^{\mathbf{x}_L(i)}}{\sum_j e^{\mathbf{x}_L(j)}}$. Inference consists of the application of this process in a feedforward manner, and training is carried out with stochastic gradient descent to minimize the cross entropy loss [4] over the training set (via back propagation).

The process of pre-training a CNN, represented by the model $\widetilde{\mathbf{y}}^* = f(\widetilde{\mathbf{x}}; \widetilde{\theta})$ (with $\widetilde{\theta} = [\widetilde{\theta}_{cn}, \widetilde{\theta}_{fc}, \widetilde{\theta}_{mn}]$), is defined as training K_1 stages of convolutional and non-linear subsampling layers (represented by the parameters $\widetilde{\theta}_{cn}$), then K_2 fully connected layers (parameters $\widetilde{\theta}_{fc}$) and one multinomial logistic regression layer $\widetilde{\theta}_{mn}$ by minimizing the cross-entropy loss function [4] over the dataset $\widetilde{\mathcal{D}}$. This pre-trained model can be used by taking the first $K_1 + K_2$ layers to initialize the training of a new model [1], in a process called fine-tuning (Fig. 2(a)). Yosinski et al. [1] notice that using a large number of pre-trained layers and fine tuning the CNN is the key to achieve the best classification results in transfer learning problems. Following this result, we take the parameters $\widetilde{\theta}_{cn}, \widetilde{\theta}_{fc}$ and add a

new multinomial logistic regression layer θ_{mn} (with random initial values), and fine tune the CNN model by minimizing the cross-entropy loss function using $\mathcal{D}$. This fine tuning process will produce six models per case: 1) MLO image $\mathbf{y} = f(\mathbf{x}_{\text{MLO}}; \theta_{\text{MLO,im}})$, 2) CC image $\mathbf{y} = f(\mathbf{x}_{\text{CC}}; \theta_{\text{CC,im}})$, 3) MLO MC map $\mathbf{y} = f(\mathbf{c}_{\text{MLO}}; \theta_{\text{MLO,mc}})$, 4) CC MC map $\mathbf{y} = f(\mathbf{c}_{\text{CC}}; \theta_{\text{CC,mc}})$, 5) MLO mass map $\mathbf{y} = f(\mathbf{m}_{\text{MLO}}; \theta_{\text{MLO,ma}})$ and 6) CC mass map $\mathbf{y} = f(\mathbf{m}_{\text{CC}}; \theta_{\text{CC,ma}})$. The final multiview training (Fig. 2(b)) takes the features from the last fully connected layer (i.e., $L-1^{th}$ layer, labeled as $\mathbf{x}_{\text{MLO},L-1}$ for the first model above, and similarly for the remaining ones) from the six models, and train a single multinomial logistic regression layer using those inputs (Fig. 2(b)), resulting in $\mathbf{y} = f(\mathbf{x}_{\text{MLO},L-1}, \mathbf{x}_{\text{CC},L-1}, \mathbf{c}_{\text{MLO},L-1}, \mathbf{c}_{\text{CC},L-1}, \mathbf{m}_{\text{MLO},L-1}, \mathbf{m}_{\text{MLO},L-1}; \theta_{mn})$, where θ_{mn} is randomly initialized in this multiview training.

3 Materials and Methods

We use the publicly available InBreast [18] and DDSM [19] mammogram datasets. InBreast [18] has 115 patients (410 images), and it does not have any division in terms of training and testing sets, so we run a 5-fold cross validation experiment, where we randomly divide the original set into 90 patients for training and validation and 25 for testing. DDSM [19] has 172 patients (680 images), obtained by joining the original MC and mass datasets proposed, but removing all cases that appear in the training set of mass and testing set of MC and vice versa. With DDSM, we run an experiment with the proposed 86 patients for training and 86 for testing. The distribution of patients in terms of BI-RADS for both datasets is depicted in Fig. 3. We use the MC and mass maps provided with these datasets, and if no mass or MC map is available for a particular case, we use a blank image of all zeros instead.

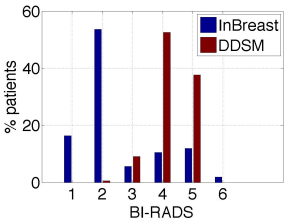


Fig. 3: Distribution of BI-RADS cases in InBreast (blue) and DDSM (red).

We use the publicly available CNN-F model from [20], consisting of a CNN that takes a 264×264 input image with four stages of convolution and non-linear sub-sampling layers, two stages of fully connected layers and a final multinomial logistic regression stage. Specifically, CNN-F has stage 1 with 64 11×11 filters and a max-pooling that sub-samples the input by 2, stage 2 with 256 5×5 filters and a max-pooling that sub-samples the input by 2, stages 3-5 with 256 3×3 filters (each) with no sub-sampling, stage 6-7 with 4096 nodes (each), and stage 8 containing the softmax layer. This model is pre-trained with Imagenet [2] (1K visual classes, 1.2M training, 50K validation and 100K test images), then we replace stage 8 with a new softmax layer containing only three classes (negative, benign and malignant) and fine tune for the CC and MLO views and the MC and mass segmentation maps (Fig. 2(a)). Finally, we take the features from stage 7 for the six models and train stage 8 of the multiview model (Fig. 2(b)). The use of the proposed pre-training can be seen as a regularization approach, which can be compared to other forms of regularization, such as data augmentation [4], obtained by artificially augmenting the training set with random geometric transformations

applied to each original sample that generates new artificial training samples. Hence, we also run an experiment that uses the same CNN-F structure with no pre-training (i.e., with random weight initialization using an unbiased Gaussian with standard deviation 0.001) and runs the training with data augmentation by adding 10 or 20 new samples per training image. Each new training sample is built by cropping the original image using randomly defined top-left and bottom-right corners (within a range of $[1, 10]$ pixels from the original corners). For completeness, we also apply this data augmentation to the pre-trained CNN-F model. Finally, the learning rate = 0.001, and momentum = 0.9. With this experimental design, we can then compare the pre-training and data augmentation regularizations.

The input CC and MLO views are pre-processed with local contrast normalization, then Otsu's segmentation [21] to remove most of the background. This pre-processing steps is found to improve the final results (Fig. 5 shows some samples of these pre-processed images). The classification accuracy is measured using volume under ROC surface (VUS) for a 3-class problem [22], and the area under ROC curve (AUC) for the benign/malignant classification of cases that have at least one finding (MC or mass).

4 Results

Figure 4 shows the VUS for the test sets of InBreast (average and standard deviation of 5-fold cross validation and the two breasts) and DDSM (average and standard deviation of the two breasts). We show the results for the six inputs per case (two views and four maps) and the result of the multiview model, and we also display the improvement achieved with the Imagenet pre-trained model compared to the randomly initialized model. Figure 5 shows four typical results of the pre-trained model (without data augmentation) on InBreast test cases. The classification of the cases into benign or malignant (where each case has at least an MC or a mass) produces an AUC of $0.91(\pm 0.05)$ on InBreast and $0.97(\pm 0.03)$ on DDSM using the pre-trained model with no data augmentation, where the same general trends can be observed compared to Fig. 4. Finally, the training time for all six models and the final multiview model (with no data augmentation) is one hour. With 10 additional training samples, the training time is four hours and with 20 additional training samples, the training time increases to 7.5 hours. These training times are obtained on InBreast, measured using Matconvnet training [20] on a 2.3GHz Intel Core i7 with 8GB, and graphics card NVIDIA GeForce GT 650M 1024 MB.

5 Discussion and Conclusion

The results show a clear improvement of multiview compared to single views, demonstrating that the high-level features of the individual CNN models provide a robust representation of the input images, which do not need to be registered. This is particularly important in mammograms, where registration is challenging due to non-rigid deformations. The poor performance of some single view classifications is due to: 1) including some MC view cases with BI-RADS > 1 that do not have annotation, which makes its classification difficult (similarly for some mass view cases); and 2) the classification using mammogram only is challenging because of the lack of sufficient detail in the image. We also notice that

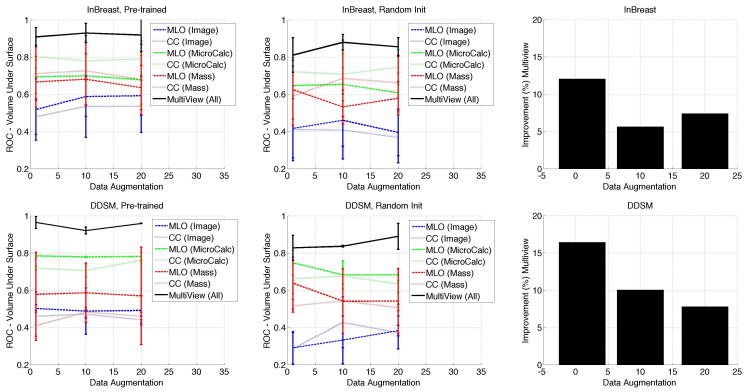


Fig. 4: VUS in terms of data augmentation on InBreast (top) DDSM (bottom). 1^{st} column shows the results with the Imagenet pre-trained model, 2^{nd} shows the randomly initialized models, and the third displays the average improvement in the results with pre-trained models, compared to the randomly initialized models.

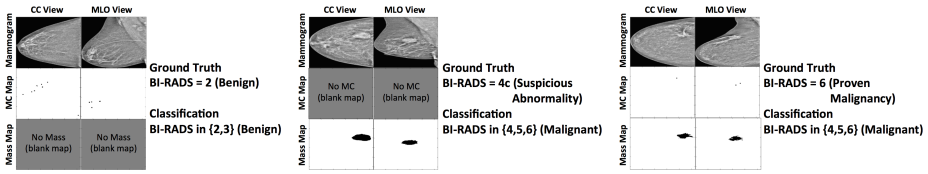


Fig. 5: InBreast test case results using Imagenet pre-trained model with no data augmentation (the ground truth and the automatic classifications are shown).

the pre-trained multiview model provides better classification results compared to the randomly initialized model, with improvements of 5% to 16%, where the largest differences happen with the no data augmentation test. This is expected given that this is the condition where the random initialized model is more likely to overfit the training data. These results are difficult to compare to previously published results in the field (see Sec. 1) given that: 1) most of the previously published results are computed with datasets that are not publicly available, and 2) these other results focus mostly on the specific classification of masses or MCs instead of the full breast exam. Nevertheless, looking exclusively at the AUC results, our methodology can be considered to be comparable (on InBreast) or superior (on DDSM) to the current state of the art, which present AUC between $[0.9, 0.95]$ for MCs and mass classification [15, 16]. However, we want to emphasize that our results can be used as baseline in the field given that we only use public data and models, and consequently be fairly compared to other works, resolving one of the issues identified by Giger et al. [14]. We believe that our work opens two research fronts that can be applied to other medical image

analysis applications, which are the use of pre-trained models from non-medical imaging datasets and the comprehensive analysis of unregistered multiview data.

References

1. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NIPS (2014) 3320–3328
2. Russakovsky, O., et al.: ImageNet Large Scale Visual Recognition Challenge (2014)
3. Bengio, Y.: Learning deep architectures for ai. Foundations and trends® in Machine Learning **2**(1) (2009) 1–127
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012)
5. Zou, W., Zhu, S., Yu, K., Ng, A.Y.: Deep learning of invariant features via simulated fixations in video. In: NIPS. (2012)
6. Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: MICCAI 2013.
7. Roth, H.R., et al.: A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In: MICCAI 2014.
8. Fakoor, R., Ladhak, F., Nazi, A., Huber, M.: Using deep learning to enhance cancer diagnosis and classification. In: ICML Workshops (2013)
9. Carneiro, G., Nascimento, J.C.: Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data. TPAMI **35**(11) (2013) 2592–2607
10. Cruz-Roa, A.A., Ovalle, J.E.A., Madabhushi, A., Osorio, F.A.G.: A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In: MICCAI 2013.
11. Li, R., Zhang, W., Suk, H.I., Wang, L., Li, J., Shen, D., Ji, S.: Deep learning based imaging data completion for improved brain disease diagnosis. In: MICCAI 2014.
12. Brosch, T., et al.: Modeling the variability in brain morphology and lesion distribution in multiple sclerosis by deep learning. In: MICCAI 2014.
13. Guo, Y., et al.: Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features. In: MICCAI 2014.
14. Giger, M.L., Karssemeijer, N., Schnabel, J.A.: Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. Annual review of biomedical engineering **15** (2013) 327–357
15. Cheng, H., Shi, X., Min, R., Hu, L., Cai, X., Du, H.: Approaches for automated detection and classification of masses in mammograms. Pattern recognition **39**(4) (2006) 646–668
16. Wei, L., Yang, Y., Nishikawa, R.M., Jiang, Y.: A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. TMI **24**(3) (2005) 371–380
17. Horsch, K., et al.: Classification of breast lesions with multimodality computer-aided diagnosis: Observer study results on an independent clinical data set. Radiology **240**(2) (2006) 357–368
18. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: Inbreast: toward a full-field digital mammographic database. Academic Radiology **19**(2) (2012) 236–248
19. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P.: The digital database for screening mammography. In: Proceedings of the 5th international workshop on digital mammography. (2000) 212–218
20. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. arXiv:1405.3531 (2014)
21. Otsu, N.: A threshold selection method from gray-level histograms. Automatica **11**(285-296) (1975) 23–27
22. Landgrebe, T.C., Duin, R.P.: Efficient multiclass roc approximation by decomposition via confusion matrix perturbation analysis. TPAMI **30**(5) (2008) 810–822