Breast Tissue Segmentation and Mammographic Risk Scoring Using Deep Learning

Kersten Petersen[1], Mads Nielsen[1,2], Pengfei Diao[1],
Nico Karssemeijer[3], and Martin Lillholm[2]

[1] Department of Computer Science, University of Copenhagen, Denmark
[2] Biomediq A/S, Copenhagen, Denmark
[3] Department of Radiology, Radboud University Nijmegen, Netherlands

Abstract. Mammographic scoring of density and texture are established methods to relate to the risk of breast cancer. We present a method that learns descriptive features from unlabeled mammograms and, using these learned features as the input to a simple classifier, address the following tasks: i) breast tissue segmentation ii) scoring of percentage mammographic density (PMD), and iii) scoring of mammographic texture (MT). Our results suggest that the learned PMD scores correlate well to manual ones, and that the learned MT scores are more related to future cancer risk than both manual and automatic PMD scores.

Keywords: Unsupervised feature learning, deep learning, breast cancer, mammograms, prognosis, risk factor, segmentation.

1 Introduction

Breast cancer is the most common cancer (non-melanoma skin cancer excluded) worldwide, with more than 430,000 deaths in 2010 alone [1]. In order to reduce breast cancer mortality, it is important to identify, monitor, and possibly treat high risk patients early. One of the strongest known risk factors for breast cancer is the relative amount of radiodense tissue in the breast, expressed as mammographic density (MD) [2][3]. Widespread MD scores range from manual categorical (e.g., BI-RADS, Wolfe [4], Tabár [5]) to continuous scores (e.g., Cumulus-like thresholding). A major problem is that fully manual or user-assisted scoring is subjective and time-consuming. There has been a trend towards fully automating MD scoring, but most of these approaches rely on handcrafted features with several adjustable hyperparameters. Similarly, mammographic texture (MT) scoring methods, describing mammographic heterogeneity, have used manually encoded and selected features.

In this paper, we investigate a method to automatically learns features that best describe mammogram appearance patterns. These data-driven features can be used to address three breast cancer risk related tasks that have previously been modeled in very different ways: breast tissue segmentation, percentual mammographic density (PMD) scoring, and mammographic texture (MT) scoring.

H. Fujita, T. Hara, and C. Muramatsu (Eds.): IWDM 2014, LNCS 8539, pp. 88–94, 2014.

2 Materials and Method

2.1 Materials

We have evaluated our method on 495 right and corresponding left mediolateral (RMLO and LMLO) mammograms from a previously published case-control study from the Dutch screening program. This study was originally designed for investigating the effect of recall rate within the Dutch biennial breast screening program [7]. Selected mammograms from this study contained 250 controls and 245 cases of which 123 were diagnosed with an interval cancer and 122 with a screen-detected cancer. The case mammograms were selected 4 years prior to a screen detected and 2-4 years prior to an interval cancer. The mammograms of the controls remained cancer free in the subsequent 4 years. The participants of the study were between 49 and 81 years old and the study groups were matched for age. The mammograms were digitized with a Vidar scanner that provided an image resolution of roughly 1500×2500 pixels on 12-bit gray scale and 50×50 microns. On the RMLO mammograms, a trained radiologist annotated the skin-air boundary and the pectoral muscle by a polygon tool, and estimated BI-RADS and PMD using a Cumulus-like approach.

2.2 Methods

The employed texture scoring method learns a deep hierarchy of increasingly more abstract features from unlabeled data and maps the final feature representation to the label of interest. Depending on the task, these per pixel labels are i) segmentation: background (BG), pectoral muscle (PM), and breast tissue (BT) ii) PMD scoring: fatty tissue, and dense tissue iii) MT scoring: healthy, and diseased (each pixel is associated with the cancer outcome label). The employed model is called a convolutional sparse autoencoder (CSAE [6]) and processes small patches at multiple image scales from the mammogram.

The training data is collected by randomly drawing 50,000 patches across a set of training mammograms and by associating them with the label of interest. An unseen mammogram is segmented or scored by applying the trained model in a sliding window approach. At the image boundary, the image is padded with a constant value. A label posterior of the disease class is gained for each location, and afterwards averaged to produce a single score per mammogram. In the following, we summarize the ingredients of the CSAE model: a convolutional architecture as the model representation, and a sparse autoencoder for learning the model parameters.

Convolutional architecture A *convolutional Architecture.* is suited for learning a deep feature hierarchy from structured data [8]. It is similar to neural networks, but has two advantages: First, convolutional architectures model the topology of the input in each layer, e.g., images as a 2D grid. Second, they are able to scale to much larger inputs by constraining the number of trainable parameters.

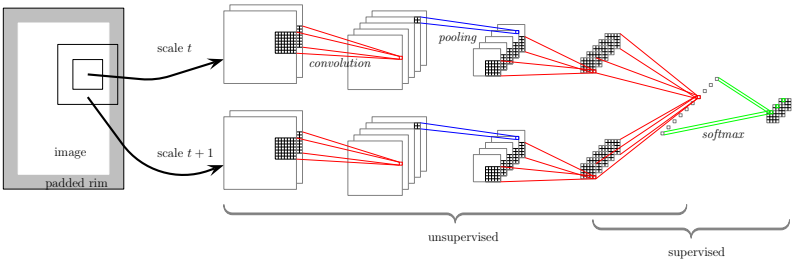

Fig. 1. Deep convolutional architecture. Patches are extracted from multiple scales of the image and fed to the convolutional architecture. The patch (or *feature map*) at scale $t+1$ only considers every second pixel, such it has the same size as the patch from scale t. The small subregions within the extracted patches are referred to as *local receptive fields*. We refer to the text for more details.

The hidden layers of a convolutional architectures consist of *convolutional* and *pooling* layers, usually in alternating order. In our convolutional architecture, we have replaced one pooling layer by a convolutional layer to have invariance to noise, but still pick up details that could benefit the segmentation or scoring task. This design choice was confirmed by slightly better results.

Convolutional layers are similar to hidden layers in a traditional neural network. They are parameterized by trainable weight parameters that can be interpreted as *features*. However, rather than connecting each unit to all input units, convolutional units are only connected to spatially close units. Each set of input units usually corresponds to a small squared subregion of the input grid and is collectively referred to as a *local receptive field*. In Figure 1, the local receptive fields are connected with red, blue, or green lines to a unit of an output feature map. Weighting the units within a local receptive field is equivalent to a convolution of the input feature map.

In a convolutional architecture, the layer units are not stored as a vector, but in a multi-channel grid structure. The convolutional layer convolves the input of each channel, sums the responses, adds a bias term, and sends the result through a scalar-wise nonlinear activation function to create one output feature map. This nonlinear multi-channel processing is repeated with different filter weights to create multiple output feature maps. Thus, each convolutional layer is fed with multiple input feature maps and applies different convolutions to create multiple output feature maps. Formally, the jth output feature map of a convolutional layer is given by

$$z_j^{\text{out}} = \sigma(w_j * z + b_j \mathbf{1}_{m'}) \,, \tag{1}$$

where w_j denotes the filter for the jth output map, z all input feature maps as a tensor, b_j the bias for the jth map, and $\mathbf{1}_{m'}$ denotes an $m' \times m'$ matrix of ones. The activation function is denoted by $\sigma(x) = \max(x, 0)$.

Table 1. Comparison of expert's PMD scores with expert's BI-RADS and automated CSAE PMD scores

Method	Case	Control	R_{PMD}	AUC (95% CI)
PMD	0.20 ± 0.13	0.18 ± 0.13	-	0.56 (0.51, 0.61)
BI-RADS	2.23 ± 0.72	2.10 ± 0.76	0.87	0.55 (0.50, 0.60)
PMD_{CSAE}	0.21 ± 0.11	0.18 ± 0.13	0.87	0.56 (0.51, 0.61)

The output of the convolutional layer is often fed to a pooling layer which summarize the distributions within small (non-overlapping) spatial regions. The final architecture layer maps the output of the last hidden layer to the labeled data, in the same way as it is modeled in a neural network.

Figure 1 illustrates our convolutional architecture for processing multiscale input patches. The large rectangles denote feature maps, whereas the small rectangles represent local receptive fields. We employ one pooling layer (blue) and three convolutional layers (red), which are trained in an unsupervised way by sparse autoencoders (see next Section). The last two layers are finally trained by a classifier to map the output to the labels of interest.

Sparse Autoencoder. Here, we describe how the w_j weights from the convolutional layer described above are trained by reconstructing the inputs of a convolution operation, i.e., patches of the size of local receptive fields, using an encoder-decoder architecture [9]. The *encoder* maps the input to a hidden layer and uses the same activation function that was defined in the convolutional layer. The *decoder* maps the hidden layer to the output layer, which is set to be the same as the input. This autoencoder structure enables to learn features without using the label information. As a refinement to this architecture, we have incorporated a sparsity regularizer to control the capacity of this model.

3 Results

3.1 Breast Tissue Segmentation

The mean and standard deviation of the Dice coefficient for automated vs. expert's breast tissue segmentation (BG: 0.99 ± 0.01, PM: 0.95 ± 0.08, and BT: 0.98 ± 0.01).

In the following, the automated breast tissue mask is used as a region of interest in both scoring tasks.

3.2 Mammographic Density Scoring

The CSAE model was trained to automatically compute PMD. Table 1 presents i) mean and standard error for cancers and controls, ii) Pearson's R correlation

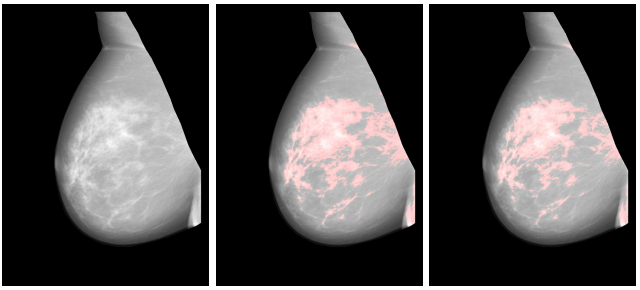

Fig. 2. Automated PMD. From let to right: original image; dense tissue (in red) based on expert Cumulus-like score; PMD_{CSAE} posterior of dense tissue class.

Table 2. Comparison of automated texture scores

Method	AUC (95% CI)
MT_{kNN} (R)	0.62 (0.57, 0.67)
MT_{CSAE} (R)	0.65 (0.60, 0.70)
MT_{CSAE} (L)	0.65 (0.60, 0.70)

coefficient between automated and manual PMD scores, and iii) the area under the ROC curve (AUC). We see that our automated PMD scores are well correlated to manual PMD and equally discriminative. A typical density segmentation result is shown in Fig. 2.

3.3 Mammographic Texture Scoring

The CSAE model has been evaluated on the LMLO and RMLO mammograms of the Nijmegen dataset. Since manual breast segmentations were only available for the RMLO view, we applied the segmentation model trained on the RMLO view to the LMLO mammograms. The obtained automated segmentations were scored with the texture model that was trained on the RMLO mammograms as well. In both experiments, the RMLO mammograms in the cross validations folds were replaced with their LMLO counterparts.

Table 2 summarizes the obtained AUCs for our model applied to RMLO, MT_{CSAE} (R), and LMLO, MT_{CSAE} (L). We also compared these models to the previously best performing MT scoring method by Nielsen et al., MT_{kNN} (R)[10]. Pearson's R correlation of the two automated MT scores to manual PMD is low (both $R_{\text{PMD}} = 0.10$), suggesting that our MT scores add to manual PMD in terms of risk segregation.

Figure 3 illustrates the correlation of the automated MT scores on LMLO and RMLO view (Pearson's $R = 0.85$), which compares to widespread volumetric density scores like Volpara$^{\text{TM}}$ ($R = 0.92$ on 2217 mammograms) [11].

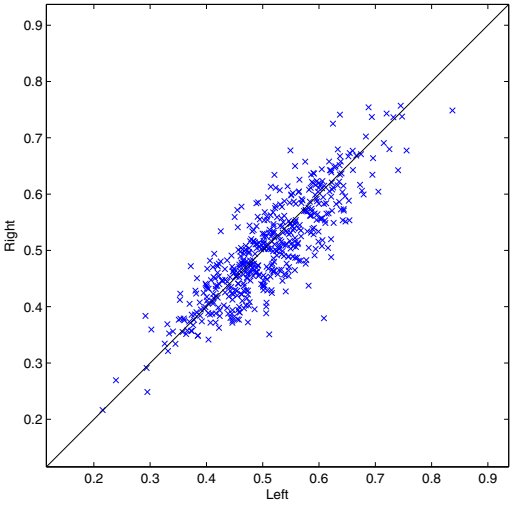

Fig. 3. Correlation of texture scores between left and right breast. The black line is the identity line. The corresponding Pearson's R correlation coefficient is 0.85.

4 Conclusion

We have presented an unsupervised feature learning method for breast region segmentation, automatic PMD scoring, and automatic MT scoring. The model learns features across multiple scales and harnesses correlations in the target values. Once the features are learned, they are fed to a simple classifier that is specific to the task of interest. The CSAE model achieved state-of-the-art results on each of the three different breast cancer related tasks.

Acknowledgements. This work was supported by the Danish National Advanced Technology Foundation under the grant Personalized Breast cancer Screening, the Danish Cancer Society, and the European Unions Seventh Framework Programme FP7 under grant agreement no 306088.

References

1. Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., Adair, T., Aggarwal, R., Ahn, S.Y., et al.: Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. The Lancet 380(9859), 2095–2128 (2013)
2. McCormack, V.A., dos Santos Silva, I.: Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis. Cancer Epidemiology Biomarkers & Prevention 15(6), 1159–1169 (2006)

3. Boyd, N.F., Martin, L.J., Bronskill, M., Yaffe, M.J., Duric, N., Minkin, S.: Breast tissue composition and susceptibility to breast cancer. Journal of the National Cancer Institute 102(16), 1224–1237 (2010)
4. Wolfe, J.N.: Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer 37(5), 2486–2492 (1976)
5. Tabár, L., Duffy, S.W., Vitak, B., Chen, H.-H., Prevost, T.C.: The natural history of breast carcinoma. Cancer 86(3), 449–462 (1999)
6. Petersen, K., Nielsen, M., Ng, A.Y., Diao, P., Vachon, C.M., Karssemeijer, N., Lillholm, M.: Unsupervised deep learning for image segmentation and mammographic risk scoring. IEEE Transactions on Medical Imaging (in review)
7. Otten, J.D., Karssemeijer, N., Hendriks, J.H., Groenewoud, J.H., Fracheboud, J., Verbeek, A.L., de Koning, H.J., Holland, R.: Effect of recall rate on earlier screen detection of breast cancers based on the dutch performance indicators. Journal of the National Cancer Institute 97(10), 748–754 (2005)
8. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
9. Ranzato, M., Poultney, C.S., Chopra, S., LeCun, Y.: Efficient learning of sparse representations with an energy-based model. In: NIPS, pp. 1137–1144 (2006)
10. Nielsen, M., Karemore, G., Loog, M., Raundahl, J., Karssemeijer, N., Otten, J., Karsdal, M., Vachon, C., Christiansen, C.: A novel and automatic mammographic texture resemblance marker is an independent risk factor for breast cancer. Cancer Epidemiology 35(4), 381–387 (2011)
11. Highnam, R., Brady, S.M., Yaffe, M.J., Karssemeijer, N., Harvey, J.: Robust breast composition measurement - volpara$^{\text{TM}}$. In: Martí, J., Oliver, A., Freixenet, J., Martí, R. (eds.) IWDM 2010. LNCS, vol. 6136, pp. 342–349. Springer, Heidelberg (2010)