Deep Multi-instance Networks with Sparse Label Assignment for Whole Mammogram Classification

Wentao Zhu(✉), Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie

Department of Computer Science, University of California, Irvine, Irvine, USA
{wentaoz1,xhx}@ics.uci.edu, {qlou,ysvang}@uci.edu

Abstract. Mammogram classification is directly related to computer-aided diagnosis of breast cancer. Traditional methods rely on regions of interest (ROIs) which require great efforts to annotate. Inspired by the success of using deep convolutional features for natural image analysis and multi-instance learning (MIL) for labeling a set of instances/patches, we propose end-to-end trained deep multi-instance networks for mass classification based on whole mammogram without the aforementioned ROIs. We explore three different schemes to construct deep multi-instance networks for whole mammogram classification. Experimental results on the INbreast dataset demonstrate the robustness of proposed networks compared to previous work using segmentation and detection annotations. (Code: https://github.com/wentaozhu/deep-mil-for-whole-mammogram-classification.git).

Keywords: Deep multi-instance learning · Whole mammogram classification · Max pooling-based MIL · Label assignment-based MIL · Sparse MIL

1 Introduction

According to the American Cancer Society, breast cancer is the most frequently diagnosed solid cancer and the second leading cause of cancer death among U.S. women [1]. Mammogram screening has been demonstrated to be an effective way for early detection and diagnosis, which can significantly decrease breast cancer mortality [15]. Traditional mammogram classification requires extra annotations such as bounding box for detection or mask ground truth for segmentation [5,11,17]. Other work have employed different deep networks to detect ROIs and obtain mass boundaries in different stages [6]. However, these methods require hand-crafted features to complement the system [12], and training data to be annotated with bounding boxes and segmentation ground truths which require expert domain knowledge and costly effort to obtain. In addition, multi-stage training cannot fully explore the power of deep networks.

Due to the high cost of annotation, we intend to perform classification based on a raw whole mammogram. Each patch of a mammogram can be treated as

M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 603–611, 2017.
DOI: 10.1007/978-3-319-66179-7_69

an instance and a whole mammogram is treated as a bag of instances. The whole mammogram classification problem can then be thought of as a standard MIL problem. Due to the great representation power of deep features [9,19–21], combining MIL with deep neural networks is an emerging topic. Yan et al. used a deep MIL to find discriminative patches for body part recognition [18]. Patch based CNN added a new layer after the last layer of deep MIL to learn the fusion model for multi-instance predictions [10]. Shen et al. employed two stage training to learn the deep multi-instance networks for pre-detected lung nodule classification [16]. The above approaches used max pooling to model the general multi-instance assumption which only considers the patch of max probability. In this paper, more effective task-related deep multi-instance models with end-to-end training are explored for whole mammogram classification. We investigate three different schemes, i.e., max pooling, label assignment, and sparsity, to perform deep MIL for the whole mammogram classification task.

The framework for our proposed end-to-end trained deep MIL for mammogram classification is shown in Fig. 1. To fully explore the power of deep MIL, we convert the traditional MIL assumption into a label assignment problem. As a mass typically composes only 2% of a whole mammogram (see Fig. 2), we further propose sparse deep MIL. The proposed deep multi-instance networks are shown to provide robust performance for whole mammogram classification on the INbreast dataset [14].

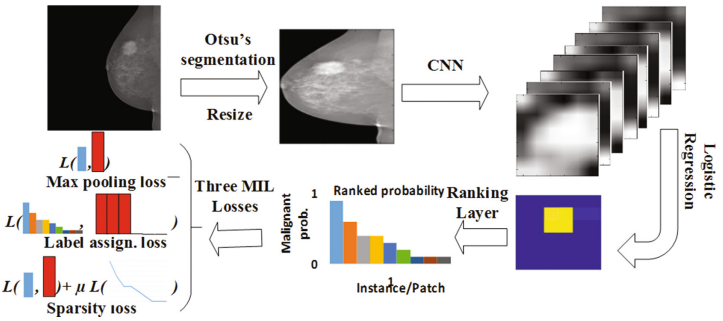
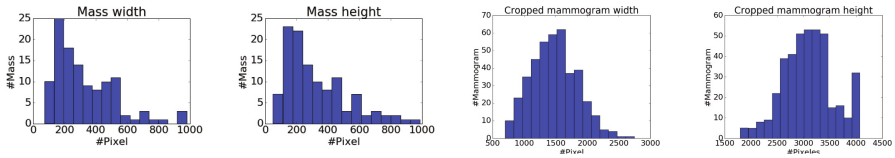


Fig. 1. The framework of whole mammogram classification. First, we use Otsu's segmentation to remove the background and resize the mammogram to 227×227. Second, the deep MIL accepts the resized mammogram as input to the convolutional layers. Here we use the convolutional layers in AlexNet [13]. Third, logistic regression with weight sharing over different patches is employed for the malignant probability of each position from the convolutional neural network (CNN) feature maps of high channel dimensions. Then the responses of the instances/patches are ranked. Lastly, the learning loss is calculated using max pooling loss, label assignment, or sparsity loss for the three different schemes.



(a) (b) (c) (d)

Fig. 2. Histograms of mass width (a) and height (b), mammogram width (c) and height (d). Compared to the size of whole mammogram ($1,474 \times 3,086$ on average after cropping), the mass of average size (329×325) is tiny, and takes about 2% of a whole mammogram.

2 Deep MIL for Whole Mammogram Mass Classification

Unlike other deep multi-instance networks [10,18], we use a CNN to efficiently obtain features of all patches (instances) at the same time. Given an image $\boldsymbol{I}$, we obtain a much smaller feature map $\boldsymbol{F}$ of multi-channels N_c after multiple convolutional layers and max pooling layers. The $(\boldsymbol{F})_{i,j,:}$ represents deep CNN features for a patch $\boldsymbol{Q}_{i,j}$ in $\boldsymbol{I}$, where i, j represents the pixel row and column index respectively, and : denotes the channel dimension.

The goal of our work is to predict whether a whole mammogram contains a malignant mass (BI-RADS $\in \{4, 5, 6\}$ as positive) or not, which is a standard binary classification problem. We add a logistic regression with weights shared across all the pixel positions following $\boldsymbol{F}$, and an element-wise sigmoid activation function is applied to the output. To clarify it, the malignant probability of feature space's pixel (i, j) is

$$r_{i,j} = \text{sigmoid}(\boldsymbol{a} \cdot \boldsymbol{F}_{i,j,:} + b), \tag{1}$$

where $\boldsymbol{a}$ is the weights in logistic regression, b is the bias, and $\cdot$ is the inner product of the two vectors $\boldsymbol{a}$ and $\boldsymbol{F}_{i,j,:}$. The $\boldsymbol{a}$ and b are shared for different pixel positions i, j. We can combine $r_{i,j}$ into a matrix $\boldsymbol{r} = (r_{i,j})$ of range $[0, 1]$ denoting the probabilities of patches being malignant masses. The $\boldsymbol{r}$ can be flattened into a one-dimensional vector as $\boldsymbol{r} = (r_1, r_2, ..., r_m)$ corresponding to flattened patches $(\boldsymbol{Q}_1, \boldsymbol{Q}_2, ..., \boldsymbol{Q}_m)$, where m is the number of patches.

2.1 Max Pooling-Based Multi-instance Learning

The general multi-instance assumption is that if there exists an instance that is positive, the bag is positive [7]. The bag is negative if and only if all instances are negative. For whole mammogram classification, the equivalent scenario is that if there exists a malignant mass, the mammogram $\boldsymbol{I}$ should be classified as positive. Likewise, negative mammogram $\boldsymbol{I}$ should not have any malignant masses. If we treat each patch $\boldsymbol{Q}_i$ of $\boldsymbol{I}$ as an instance, the whole mammogram classification is a standard multi-instance task.

For negative mammograms, we expect all the r_i to be close to 0. For positive mammograms, at least one r_i should be close to 1. Thus, it is natural to use the maximum component of $\boldsymbol{r}$ as the malignant probability of the mammogram $\boldsymbol{I}$

$$p(y = 1|\boldsymbol{I}, \boldsymbol{\theta}) = \max\{r_1, r_2, ..., r_m\}, \tag{2}$$

where $\boldsymbol{\theta}$ is the weights in deep networks.

If we sort $\boldsymbol{r}$ first in descending order as illustrated in Fig. 1, the malignant probability of the whole mammogram $\boldsymbol{I}$ is the first element of ranked $\boldsymbol{r}$ as

$$\begin{aligned} &\{r'_1, r'_2, ..., r'_m\} = \text{sort}(\{r_1, r_2, ..., r_m\}), \\ &p(y = 1|\boldsymbol{I}, \boldsymbol{\theta}) = r'_1, \quad \text{and} \quad p(y = 0|\boldsymbol{I}, \boldsymbol{\theta}) = 1 - r'_1, \end{aligned} \tag{3}$$

where $\boldsymbol{r}' = (r'_1, r'_2, ..., r'_m)$ is descending ranked $\boldsymbol{r}$. The cross entropy-based cost function can be defined as

$$\mathcal{L}_{maxpooling} = -\frac{1}{N}\sum_{n=1}^{N} \log(p(y_n|\boldsymbol{I}_n, \boldsymbol{\theta})) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \tag{4}$$

where N is the total number of mammograms, $y_n \in \{0, 1\}$ is the true label of malignancy for mammogram $\boldsymbol{I}_n$, and λ is the regularizer that controls model complexity.

One disadvantage of max pooling-based MIL is that it only considers the patch $\boldsymbol{Q}'_1$ (patch of the max malignant probability), and does not exploit information from other patches. A more powerful framework should add task-related priori, such as sparsity of mass in whole mammogram, into the general multi-instance assumption and explore more patches for training.

2.2 Label Assignment-Based Multi-instance Learning

For the conventional classification tasks, we assign a label to each data point. In the MIL scheme, if we consider each instance (patch) $\boldsymbol{Q}_i$ as a data point for classification, we can convert the multi-instance learning problem into a label assignment problem.

After we rank the malignant probabilities $\boldsymbol{r} = (r_1, r_2, ..., r_m)$ for all the instances (patches) in a whole mammogram $\boldsymbol{I}$ using the first equation in Eq. 3, the first few r'_i should be consistent with the label of whole mammogram as previously mentioned, while the remaining patches (instances) should be negative. Instead of adopting the general MIL assumption that only considers the $\boldsymbol{Q}'_1$ (patch of malignant probability r'_1), we assume that (1) patches of the first k largest malignant probabilities $\{r'_1, r'_2, ..., r'_k\}$ should be assigned with the same class label as that of whole mammogram, and (2) the rest patches should be labeled as negative in the label assignment-based MIL.

After the ranking/sorting layer using the first equation in Eq. 3, we can obtain the malignant probability for each patch

$$p(y = 1|\boldsymbol{Q}'_i, \boldsymbol{\theta}) = r'_i, \quad \text{and} \quad p(y = 0|\boldsymbol{Q}'_i, \boldsymbol{\theta}) = 1 - r'_i. \tag{5}$$

The cross entropy loss function of the label assignment-based MIL can be defined

$$\mathcal{L}_{labelassign.} = -\frac{1}{mN}\sum_{n=1}^{N}\Bigg(\sum_{j=1}^{k}\log(p(y_n|\boldsymbol{Q'}_j,\boldsymbol{\theta}))+ \sum_{j=k+1}^{m}\log(p(y=0|\boldsymbol{Q'}_j,\boldsymbol{\theta}))\Bigg)+\frac{\lambda}{2}\|\boldsymbol{\theta}\|^2. \tag{6}$$

One advantage of the label assignment-based MIL is that it explores all the patches to train the model. Essentially it acts a kind of data augmentation which is an effective technique to train deep networks when the training data is scarce. From the sparsity perspective, the optimization problem of label assignment-based MIL is exactly a k-sparse problem for the positive data points, where we expect $\{r'_1, r'_2, ..., r'_k\}$ being 1 and $\{r'_{k+1}, r'_{k+2}, ..., r'_m\}$ being 0. The disadvantage of label assignment-based MIL is that it is hard to estimate the hyper-parameter k. Thus, a relaxed assumption for the MIL or an adaptive way to estimate the hyper-parameter k is preferred.

2.3 Sparse Multi-instance Learning

From the mass distribution, the mass typically comprises about 2% of the whole mammogram on average (Fig. 2), which means the mass region is quite sparse in the whole mammogram. It is straightforward to convert the mass sparsity to the malignant mass sparsity, which implies that $\{r'_1, r'_2, ..., r'_m\}$ is sparse in the whole mammogram classification problem. The sparsity constraint means we expect the malignant probability of part patches r'_i being 0 or close to 0, which is equivalent to the second assumption in the label assignment-based MIL. Analogously, we expect r'_1 to be indicative of the true label of mammogram $\boldsymbol{I}$.

After the above discussion, the loss function of sparse MIL problem can be defined

$$\mathcal{L}_{sparse} = \frac{1}{N}\sum_{n=1}^{N}\big(-\log(p(y_n|\boldsymbol{I}_n,\boldsymbol{\theta})) + \mu\|\boldsymbol{r}'_n\|_1\big) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2, \tag{7}$$

where $p(y_n|\boldsymbol{I}_n,\boldsymbol{\theta})$ can be calculated in Eq. 3, $\boldsymbol{r}_n = (r'_1, r'_2, ..., r'_m)$ for mammogram $\boldsymbol{I}_n$, $\|\cdot\|_1$ denotes the $\mathcal{L}_1$ norm, μ is the sparsity factor, which is a trade-off between the sparsity assumption and the importance of patch $\boldsymbol{Q'}_1$.

From the discussion of label assignment-based MIL, this learning is a kind of exact k-sparse problem which can be converted to $\mathcal{L}_1$ constrain. One advantage of sparse MIL over label assignment-based MIL is that it does not require assign label for each patch which is hard to do for patches where probabilities are not too large or small. The sparse MIL considers the overall statistical property of $\boldsymbol{r}$.

Another advantage of sparse MIL is that, it has different weights for general MIL assumption (the first part loss) and label distribution within mammogram (the second part loss), which can be considered as a trade-off between max pooling-based MIL (slack assumption) and label assignment-based MIL (hard assumption).

3 Experiments

We validate the proposed models on the most frequently used mammographic mass classification dataset, INbreast dataset [14], as the mammograms in other datasets, such as DDSM dataset [4], are of low quality. The INbreast dataset contains 410 mammograms of which 100 containing malignant masses. These 100 mammograms with malignant masses are defined as positive. For fair comparison, we also use 5-fold cross validation to evaluate model performance as [6]. For each testing fold, we use three folds for training, and one fold for validation to tune hyper-parameters. The performance is reported as the average of five testing results obtained from cross-validation.

We employ techniques to augment our data. For each training epoch, we randomly flip the mammograms horizontally, shift within 0.1 proportion of mammograms horizontally and vertically, rotate within 45 degree, and set 50×50 square box as 0. In experiments, the data augmentation is essential for us to train the deep networks.

For the CNN network structure, we use AlexNet and remove the fully connected layers [13]. Through CNN, the mammogram of size 227×227 becomes 256 6×6 feature maps. Then we use steps in Sect. 2 to do MIL. Here we employ weights pretrained on the ImageNet due to the scarce of data. We use Adam optimization with learning rate 5×10^{-5} for training models [2]. The λ for max pooling-based and label assignment-based MIL are 1×10^{-5}. The λ and μ for sparse MIL are 5×10^{-6} and 1×10^{-5} respectively. For the label assignment-based MIL, we select k from $\{1, 2, 4, 6, 8\}$ based on the validation set.

We firstly compare our methods to previous models validated on DDSM dataset and INbreast dataset in Table 1. Previous hand-crafted feature-based methods require manually annotated detection bounding box or segmentation ground truth even in test denoting as manual [3,8,17]. The feat. denotes requiring hand-crafted features. Pretrained CNN uses two CNNs to detect the mass region and segment the mass, followed by a third CNN to do mass classification on the detected ROI region, which requires hand-crafted features to pretrain the

Table 1. Accuracy comparisons of the proposed deep MILs and related methods on test sets.

Methodology	Dataset	Set-up	Accu	AUC
Ball et al. [3]	DDSM	Manual+feat	0.87	N/A
Varela et al. [17]	DDSM	Manual+feat	0.81	N/A
Domingues et al. [8]	INbr	Manual+feat	0.89	N/A
Pretrained CNN [6]	INbr	Auto.+feat	0.84 ±0.04	0.69 ±0.10
Pretrained CNN+Random Forest [6]	INbr	Auto.+feat	**0.91 ± 0.02**	0.76 ±0.23
AlexNet	INbr	Auto	0.81 ±0.02	0.79 ±0.03
AlexNet+Max Pooling MIL	INbr	Auto	0.85 ±0.03	0.83 ±0.05
AlexNet+Label Assign. MIL	INbr	Auto	0.86 ±0.02	0.84 ±0.04
AlexNet+Sparse MIL	INbr	Auto	0.90 ±0.02	**0.89 ± 0.04**

network and needs multi-stages training [6]. Pretrained CNN+Random Forest further employs random forest and obtained 7% improvement. These methods are either manually or need hand-crafted features or multi-stages training, while our methods are totally automated, do not require hand-crafted features or extra annotations even on training set, and can be trained in an end-to-end manner.

The max pooling-based deep MIL obtains better performance than the pre-trained CNN using 3 different CNNs and detection/segmentation annotation in the training set. This shows the superiority of our end-to-end trained deep MIL for whole mammogram classification. According to the accuracy metric, the sparse deep MIL is better than the label assignment-based MIL, which is better than the max pooling-based MIL. This result is consistent with previous discussion that the sparsity assumption benefited from not having hard constraints of the label assignment assumption, which employs all the patches and is more efficient than max pooling assumption. Our sparse deep MIL achieves competitive accuracy to random forest-based pretrained CNN, while much higher AUC than

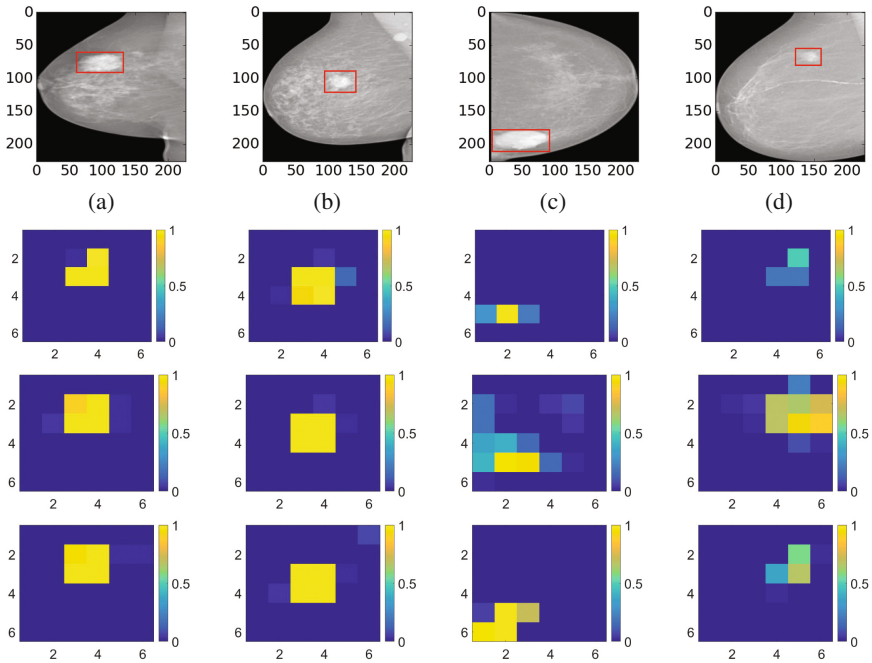


Fig. 3. The visualization of predicted malignant probabilities for instances/patches in four resized mammograms. The first row is the resized mammogram. The red rectangle boxes are mass regions from the annotations on the dataset. The color images from the second row to the last row are the predicted malignant probability from logistic regression layer for (a) to (d) respectively, which are the malignant probabilities of patches/instances. Max pooling-based, label assignment-based, sparse deep MIL are in the second row, third row, fourth row respectively.

previous work, which shows our method is more robust. The main reasons for the robust results using our models are as follows. Firstly, data augmentation is an important technique to increase scarce training datasets and proves useful here. Secondly, the transfer learning that employs the pretrained weights from ImageNet is effective for the INBreast dataset. Thirdly, our models fully explore all the patches to train our deep networks thereby eliminating any possibility of overlooking malignant patches by only considering a subset of patches. This is a distinct advantage over previous networks that employ several stages consisting of detection and segmentation.

To further understand our deep MIL, we visualize the responses of logistic regression layer for four mammograms on test set, which represents the malignant probability of each patch, in Fig. 3. We can see the deep MIL learns not only the prediction of whole mammogram, but also the prediction of malignant patches within the whole mammogram. Our models are able to learn the mass region of the whole mammogram without any explicit bounding box or segmentation ground truth annotation of training data. The max pooling-based deep multi-instance network misses some malignant patches in (a), (c) and (d). The possible reason is that it only considers the patch of max malignant probability in training and the model is not well learned for all patches. The label assignment-based deep MIL mis-classifies some patches in (d). The possible reason is that the model sets a constant k for all the mammograms, which causes some misclassifications for small masses. One of the potential applications of our work is that these deep MIL networks could be used to do weak mass annotation automatically, which provides evidence for the diagnosis.

4 Conclusion

In this paper, we propose end-to-end trained deep MIL for whole mammogram classification. Different from previous work using segmentation or detection annotations, we conduct mass classification based on whole mammogram directly. We convert the general MIL assumption to label assignment problem after ranking. Due to the sparsity of masses, sparse MIL is used for whole mammogram classification. Experimental results demonstrate more robust performance than previous work even without detection or segmentation annotation in the training.

In future work, we plan to extend the current work by: (1) incorporating multi-scale modeling such as spatial pyramid to further improve whole mammogram classification, (2) employing the deep MIL to do annotation or provide potential malignant patches to assist diagnoses, and (3) applying to large datasets and expected to have improvement if the big dataset is available.

References

1. American cancer society. what are the key statistics about breast cancer?
2. Ba, J., Kingma, D.: Adam: a method for stochastic optimization. In: ICLR (2015)

3. Ball, J.E., Bruce, L.M.: Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation. In: EMBS (2007)
4. Bowyer, K., Kopans, D., Kegelmeyer, W., et al.: The digital database for screening mammography. In: IWDM (1996)
5. Carneiro, G., Nascimento, J., Bradley, A.P.: Unregistered multiview mammogram analysis with pre-trained deep learning models. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 652–660. Springer, Cham (2015). doi:10.1007/978-3-319-24574-4_78
6. Dhungel, N., Carneiro, G., Bradley, A.P.: The automated learning of deep features for breast mass classification from mammograms. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 106–114. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_13
7. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1), 31–71 (1997)
8. Domingues, I., Sales, E., Cardoso, J., Pereira, W.: Inbreast-database masses characterization. In: XXIII CBEB (2012)
9. Greenspan, H., van Ginneken, B., Summers, R.M.: Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE TMI **35**(5), 1153–1159 (2016)
10. Hou, L., Samaras, D., Kurc, T.M., et al.: Patch-based convolutional neural network for whole slide tissue image classification arXiv:1504.07947 (2015)
11. Jiao, Z., Gao, X., Wang, Y., Li, J.: A deep feature based framework for breast masses classification. Neurocomputing **197**, 221–231 (2016)
12. Kooi, T., Litjens, G., van Ginneken, B., et al.: Large scale deep learning for computer aided detection of mammographic lesions. Med. Image Anal. **35**, 303–312 (2017)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
14. Moreira, I.C., Amaral, I., Domingues, I., et al.: Inbreast: toward a full-field digital mammographic database. Academic radiology (2012)
15. Oeffinger, K.C., Fontham, E.T., Etzioni, R., et al.: Breast cancer screening for women at average risk: 2015 guideline update from the American cancer society. Jama (2015)
16. Shen, W., Zhou, M., Yang, F., Dong, D., Yang, C., Zang, Y., Tian, J.: Learning from experts: developing transferable deep features for patient-level lung cancer prediction. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 124–131. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_15
17. Varela, C., Timp, S., Karssemeijer, N.: Use of border information in the classification of mammographic masses. Phys. Med. Biol. **51**(2), 425 (2006)
18. Yan, Z., Zhan, Y., Peng, Z., et al.: Multi-instance deep learning: discover discriminative local anatomies for bodypart recognition. IEEE Trans. Med. Imaging **35**(5), 1332–1343 (2016)
19. Zhu, W., Lan, C., Xing, J., et al.: Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In: AAAI (2016)
20. Zhu, W., Miao, J., Qing, L., Huang, G.B.: Hierarchical extreme learning machine for unsupervised representation learning. In: IJCNN, pp. 1–8. IEEE (2015)
21. Zhu, W., Xie, X.: Adversarial deep structural networks for mammographic mass segmentation arXiv:1612.05970 (2016)