Eurographics Workshop on Visual Computing for Biology and Medicine (2017)
I. Hotz, D. Merhof, and C. Rieder (Editors)

Short Paper

Maximizing AUC with Deep Learning for Classification of Imbalanced Mammogram Datasets

J. Sulam[1]†, R. Ben-Ari[2] and P. Kisilev[3]‡

[1]Computer Science Department, Technion - Israel Institue of Technology
[2]IBM Research, Haifa, Israel
[3]Huawei, Israel

Abstract

Breast cancer is the second most common cause of death in women. Computer-aided diagnosis typically demand for carefully annotated data, precise tumor allocation and delineation of the boundaries, which is rarely available in the medical system. In this paper we present a new deep learning approach for classification of mammograms that requires only a global binary label. Traditional deep learning methods typically employ classification error losses, which are highly biased by class imbalance – a situation that naturally arises in medical classification problems. We hereby suggest a novel loss measure that directly maximizes the Area Under the ROC Curve (AUC), providing an unbiased loss. We validate the proposed model on two mammogram datasets: IMG, comprising of 796 patients, 80 positive (164 images) and 716 negative (1869 images), and the publicly available dataset INbreast. Our results are encouraging, as the proposed scheme achieves an AUC of 0.76 and 0.65 for IMG and INbreast, respectively.

1. Introduction

Breast cancer is one of the most commonly diagnosed forms of cancer among women in the world [BL08], and mammographic examination constitute the most basic type of screening for this disease. As such, there is a need for reliable automatic or computer-aided diagnostic (CAD) systems. Most approaches rely on classic computer vision and classification tools, and consist of a two-stage process: an initial detection of potential abnormal candidates, and their posterior classification as malignant or benign. These methods rely on finely annotated data requiring the location and often segmentation of the tumor, which implies expensive and tedious labor from expert radiologists. Such detailed annotations are rarely available in practice, limiting their applicability. In this work we study the problem of classification of globally-labeled mammograms, without any local annotations. This is a very challenging task, as often the size of the lesion can be orders of magnitude smaller than the image, as shown in Figure 1.

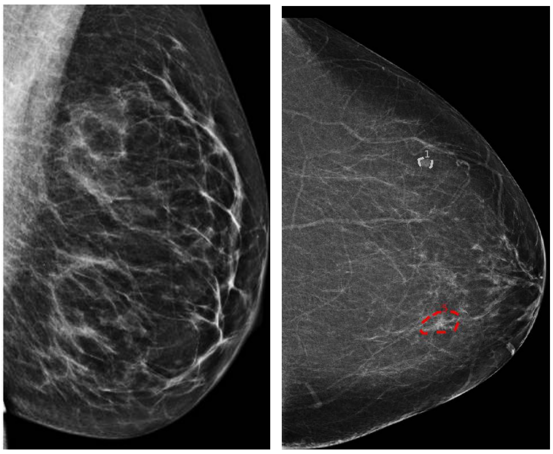

Figure 1: *Left: normal mammogram. Right: mammogram with a malignant tumor, which has been annotated in red for illustration purposes. Note the small size of the finding, yet determining the class of the whole image as positive (malignant finding).*

Data for binary classification often exhibits a highly skewed class distribution, i.e. most samples belong to a majority class. In the medical domain this scenario arises naturally as the number of healthy (normal) cases is commonly orders of magnitude higher than ill (positive) counterparts. Most learning methods minimize a classification loss based on classification accuracy – a metric that is highly biased by the class population. Strategies such as class re-sampling or data augmentation are often employed to somewhat mitigate this issue [HLCLT16], but these approaches are far from optimal. Class re-sampling and data augmentation also introduce further hyper-parameters, which makes these approaches more intricate in practice. The Area Under the ROC Curve (AUC), on the other hand, is insensitive to class distributions [CM03]. In fact,

† Research conducted as an intern at IBM Research - Haifa.
‡ Research conducted while at IBM Research - Haifa.

as the class imbalance grows the solution achieved by algorithms maximizing the AUC is increasingly better than those maximizing classification accuracy [AS05].

In this work, and unlike most current CAD methods, we address the very challenging problem of mammogram classification without any local annotations of the tumor and employing only a global label per screening - what we refer to as global labeling. Our work entails the following contributions:

1. We suggest a new convolutional neural network (CNN) architecture combining a pre-trained model and a customized CNN for classification of mammograms requiring only a global label for training.
2. The network architecture is simple and easily configurable.
3. The proposed deep convolutional network is trained with a new formulation of a loss function that maximizes the AUC measure.
4. We demonstrate our approach on the publicly available INbreast dataset and on proprietary multi-center database, IMG.

Our results indicate not only that considerable classification accuracy can be obtained in this very challenging setting, but also show the benefits of the proposed AUC-maximizing framework in the deep learning setting.

2. Current CAD Approaches for Breast Cancer

Breast cancer produces a series of organic changes in the breast tissue, which manifest as micro-calcifications, masses, architectural distortion and asymmetry. Most CAD systems focus on detecting and further classifying these different abnormalities [LCCM16], which often relies on the design of ad-hoc and handcrafted features. For instance, a traditional (not convolutional) neural network was employed in [PBP*08] to detect micro-calcifications in digitized mammograms. The work in [JY15] proposed an algorithm for the automatic detection of abnormal mammograms using features based on low level computer vision and Principal Component Analysis (PCA). The reader is referred to [LCCM16] for a more thorough review. Importantly, all these approaches require the ground-truth local annotation of the tumor, making them unsuitable for screening with only global labels.

A number of methods have been recently developed deploying deep learning techniques for mammogram classification. The work in [LJ16] employed CNN to classify pre-segmented masses in mammographies. Similar approaches were proposed in [DCB16, JGWL16]. These methods, naturally, are unable to manage the setting we tackle in this work, as they only classify pre-segmented masses. An exception is the work of [HK16], which proposed a Self-Transfer Learning framework for the classification and localization of lesions in weakly labeled mammograms. Their solution is given in terms of two networks, a classifier and a localization network. Our work is different in that we concentrate in the design and training formulation of the classification network – which could further benefit approaches such that in [HK16]. We will refer to this work again in the Results section.

3. AUC Maximization for Deep Convolutional Networks

The design of classifiers to maximize the AUC has received increasing attention in recent years. The work in [HR04] introduced a gradient descent algorithm to optimize a linear classifier with an AUC-driven loss, and on-line variants have also been proposed [ZJYH11, GJZZ13]. These methods, however, considered either linear classifiers or SVMs [AS05]. Very few attempts have been made to address this problem in convolutional deep networks. The work in [CB12] showed that maximizing the AUC in a multilayer perceptron network with a simple loss function provided better results in small and imbalanced dataset. The recent work of [WSX16] proposed an AUC-driven loss for deep belief networks, although for the case of structured data as in protein sequencing.

Convolutional neural networks that implement complex nonlinear functions can be expressed as the composition of simpler blocks, where each l^{th} layer is parameterized by weights $\mathbf{w}_l$, receiving the input $\mathbf{x}_l$ and producing the input to the following layer, $\mathbf{x}_{l+1} = f_l(\mathbf{w}_l, \mathbf{x}_l)$. Considering a CNN with L layers for binary classification, the last layer typically computes the score (or probability) of the input image $\mathbf{x}$, denoted by x_L. The training of the model reduces to minimizing a loss function that will encourage these scores to be similar to their respective labels by back-propagation. Typically, this loss is just the ℓ_2-difference between the label and the obtained probability, whereas in this work we provide a cost function that maximizes the AUC.

Consider the scenario of a set of N^+ training examples $\mathbf{x}_i$ from a positive (minority) class, and N^- examples $\mathbf{x}_j$ from a negative (majority) class, i.e. $N^- > N^+$, each with label y_i. Given a classifier f, the AUC can be estimated through the Mann-Whitney statistic [MW47], as

$$\text{AUC} = \frac{1}{N^+N^-}\sum_{i=1}^{N^+}\sum_{j=1}^{N^i}\mathbb{I}_{(f(\mathbf{x}_i)>f(\mathbf{x}_j))}, \tag{1}$$

where $\mathbb{I}$ is the indicator function. One could indeed maximize the this measure by proposing a loss function that penalizes cases where a positive sample $\mathbf{x}_i$ is assigned a lower or equal score than a negative one $\mathbf{x}_j$, i.e. when $f(\mathbf{x}_i) \leq f(\mathbf{x}_j)$. We define such loss by means of an appropriate function ℓ, while adding a regularization term on the network weights, symbolically represented by $\mathbf{W}$, i.e.

$$\mathcal{L}(\mathbf{W}) = \frac{\lambda}{2}\|\mathbf{W}\|_2^2 + \frac{1}{N^+N^-}\sum_{i=1}^{N^+}\sum_{j=1}^{N^i}\ell(\mathbf{x}^i,\mathbf{x}^j). \tag{2}$$

Moving to an online optimization scheme, the above loss can be modified to become the sum of a loss over individual samples during training at time t, i.e., $\mathcal{L}(\mathbf{W}) = \sum_{t=1}^{T}\mathcal{L}_t(\mathbf{W})$. In this setting, the computed loss for every training sample $(\mathbf{x}^t, y^t)$ is given by

$$\begin{aligned}\mathcal{L}_t(x_L^t, y^t) = \frac{\lambda}{2}\|\mathbf{W}\|_2^2 + \mathbb{I}_{(y^t=1)}\frac{1}{N^-}\sum_{j=1}^{N^-}\ell(x_L^t, x_L^j)\\ +\mathbb{I}_{(y^t=-1)}\frac{1}{N^+}\sum_{i=1}^{N^+}\ell(x_L^i, x_L^t).\end{aligned} \tag{3}$$

When the above $\ell(x_L^t, x_L^j)$ is a step function of the difference $(x_L^j - x_L^t)$, the expression in Eq. (3) effectively maximizes an online estimation of the AUC [ZJYH11] (plus the regularazation term). One should propose a smooth surrogate function for the above term, and we define it as a variant of the logistic function,

$$\ell(x^i, x^j) = -\log\left(\frac{1}{1+e^{(x^j-x^i)}}\right). \tag{4}$$

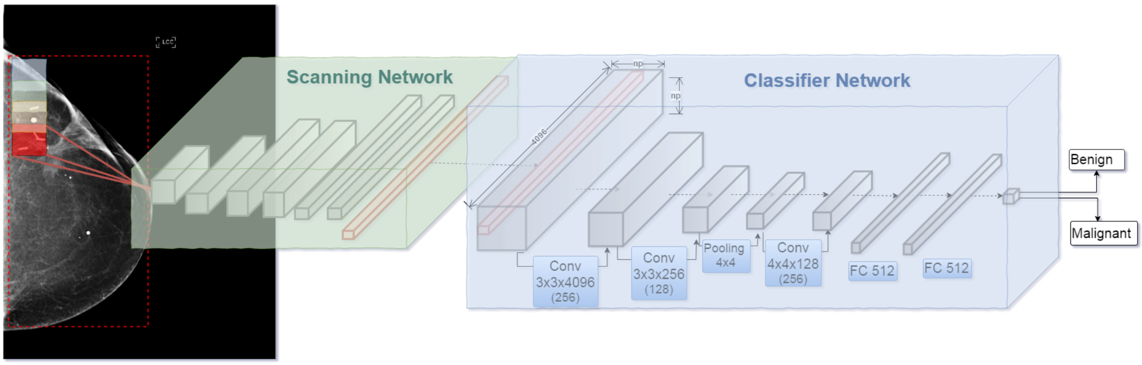

Figure 2: *The complete architecture, comprised of a first scanning network and a classifier network.*

Note that to employ back propagation, one only needs to compute $\partial\mathcal{L}_t/\partial x^t_L$, which reduces to the sum of the terms $\frac{\partial \ell}{\partial x^t_L}$. These partial derivatives are very simple to compute as they are nothing but the derivatives of a shifted logistic function.

The difficulty in the minimization of Eq. (3) resides in the pairwise nature of the loss between samples of different classes: each sample $\mathbf{x}^t$ also depends on other samples $\mathbf{x}^i$ from the opposite class. This might be problematic not only because of the added computational complexity, but also because keeping all previous samples in memory is often infeasible. Interestingly, and unlike previous approaches [ZJYH11, GJZZ13], the latter is not a problem in the deep learning framework, as one only needs to compute the gradient of the loss w.r.t x^i_L – and not w.r.t the high dimensional images $\mathbf{x}^i$. Since these elements x^i_L are scalars (simply the values assigned to each previously seen sample), they can be easily stored in memory. To this end, we propose to keep two buffers of positive and negative classes, $\mathcal{S}^+ = \{x^i_L : y^i = 1\}$ and $\mathcal{S}^- = \{x^i_L : y^i = -1\}$, with which to compute Eq. (3). From a complexity perspective, evaluating $\mathcal{L}_t(x^t_L)$ is $\mathcal{O}(N^2)$. In the context of mammography classification, where the number of training examples is in the order of a few thousands, this does not constitute a limitation. However, this simple approach might become prohibitive when considering big-data scenarios, where the size of the dataset grows considerably. In such cases one can easily modify the scheme above by keeping only a (properly) sub-sampled version of the buffers $\mathcal{S}^+$ and $\mathcal{S}^-$ as done in [ZJYH11], reducing the complexity to $\mathcal{O}(N)$.

4. The proposed architecture

The problem of classifying a mammogram screening based only on a single label is a challenging task, though very relevant in practice. Mammographic images are very high-dimensional, usually in the order of 1K×3K pixels – an order of magnitude larger than common natural images benchmarks. On the other hand, while the size of the image is large, malignant findings can be as small as tens of pixels in width and height, as exemplified in Figure 1. A naive approach would be to resize the images to a standard 224×244 image so as to employ off-the-shelf CNN models that require this input size. However, this severe shrinking inevitably causes loss of information and details, which might be critical for the classification process. Thus, we present a network architecture that processes the mammogram in full resolution in order to capture even small findings that can potentially determine the label of the whole image. To this end, we decompose the image into a grid of overlapping patches that are analyzed with a screening network, generating a global representation for the input mammogram. This representation is subsequently analyzed by a classifier CNN that effectively assigns a global label to the input. This general scheme is depicted in Figure 2.

The scanning CNN constitutes the first step of the analysis of the mammogram, and it is a completely convolutional network whose main task is to provide feature descriptors of local areas in the mammogram. Training such a network from scratch might be problematic due to insufficient data, so we leverage the representation power of the popular VGG-m network [CSVZ14] and employ its first 5 convolutional layers for this stage, disregarding its last two fully-connected layers. This choice is motivated by choosing a powerful feature extraction architecture while disregarding the last classification-focused layers, as it is common in Transfer Learning. Because of its convolutional properties, this first network scans the entire image producing a feature vector of fixed size per location. In practice we restrict the effective area of the image to an automatically determined bounding box. Each window produces a 4096 features long vector, and n_p overlapping local windows are analyzed in each vertical and horizontal directions ($n_p = 15$ in our experiments). These features, which represent and characterize a local neighborhood are then grouped together preserving their spatial relation, resulting in a representation volume of $n_p \times n_p \times 4096$. Note that this particular aggregation strategy guarantees the spatial proximity of features originated from proximal regions.

The output of this stage is fed into a second network, whose objective is to classify the obtained representation as containing or not a malignant lesion. To this end, we propose an architecture that combines convolutions, REctifying Linear Units (RELUs) and max-pooling operations, in order to decrease the spatial dimensions to a single pixel. As detailed in Figure 2, we employ 256 filters in the first layer, 128 in the second layer and 256 in the third convolutional layers. Finally, two fully connected (FC) layers of 512 filters each perform the final classification yielding a singular output, corresponding to the probability of presence or absence of a malignant lesion. Certainly, other choices of architectures are possible and designing better alternatives is subject of ongoing work.

Recall that while these two networks can be explained separately,

Table 1: *Results of the proposed approach, as well as other competing methods.*

Method	dataset	Algorithm Input	Labels	AUC
Dhungel et al [DCB16]	INbreast	Segmented Masses	Birad +masses bounding box	0.76
Dhungel et al [DCB17]	INbreast	Mammograms (joint MLO+CC)	Global Label Birad	0.74
Baseline 1 (this work)	INbreast	Mammograms (resized to 224×224)	Global Label Birad	0.602
Baseline 2 (this work)	INbreast	Mammograms (original scale)	Global Label Birad	0.630
Proposed Method	INbreast	Mammograms (original scale)	Global Label Birad	**0.650**
Baseline 1 (this work)	Proprietary	Mammograms (resized to 224×224)	Global Label Birad	0.677
Baseline 2 (this work)	Proprietary	Mammograms (original scale)	Global Label Birad	0.727
Proposed Method	Proprietary	Mammograms (original scale)	Global Label Birad	**0.767**

they can also be understood as an integrated complete deep convolutional neural network. Indeed, both architectures can be merged together by setting the stride of the first layer of the scanning network accordingly. The training of the model consists mainly on training the classification network (while possibly fine-tuning the first scanning stage), given the positive and negative image labels. In practice, and to speed-up training, one can pre-compute the volumetric representations of all images in the training set. After a certain number of training epochs, and given sufficient training data, one can merge both networks back together and continue the training further, improving performance while fine-tuning.

5. Experiments and Results

We validate the proposed model on two dataset: the first one, IMG, is a proprietary mammogram dataset comprising of 796 patients, 80 of them defined as positive (164 images BIRAD ≥ 4), and 716 negative (1869 images) with both Cranial-Caudal (CC) and Mediolateral-Oblique (MLO) views, belonging to normal patients as well as benign findings (BIRADS ≤ 2), resulting from a large multi-center study and examined by expert radiologists. The second dataset is the publicly available INbreast dataset [MAD*12], consisting of 115 cases with 410 images. We define positive and negatives classes in the same manner as for the previous dataset. The splitting of positive and negative classes follows the rationale of distinguishing between severe abnormalities from normal images (BIRADS 1) and most-likely-benign findings (BI-RADS 2). This splitting makes the classification task all the more challenging, as it is not enough to detect abnormalities (e.g., masses) but the system must also discriminate benign (rounded) from malignant ones (spiculated).

We compare our method with 2 baseline algorithms. Baseline 1 comprises of a naive transfer learning strategy, in which the image (after cropping the area of the breast) is resized to a size of 224×224 pixels. The breast image is then run through an off-the-shelf deep CNN model, trained on the Imagenet dataset (we employ the same VGG-m model just as described before), obtaining from it a 4096 long representation vector for the entire image. These features are then used to train an SVM in order to classify them as positive or negatives. Note that this scheme, while it benefits from a similar feature-extraction machinery, does not have the advantages of analyzing local or detailed structures. The second baseline consists of the very same deep CNN architecture described in Figure 2, though the learning of the network is done with a traditional logistic (binary) loss. This scheme benefits from the same analysis power of the proposed approach, but does not leverage our AUC-maximization formulation.

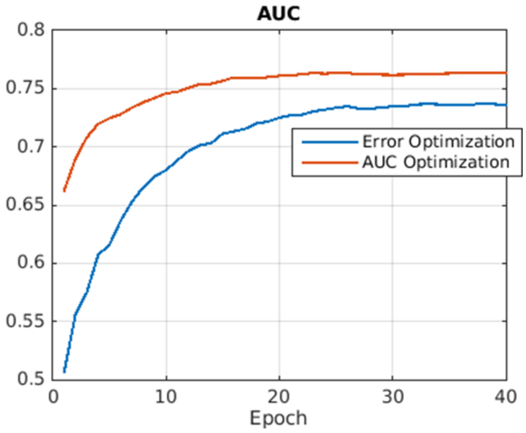

Figure 3: *Evolution of the AUC with the iterations, for one of the splits in the proprietary dataset, for the AUC optimization framework compared to the traditional Error-Loss.*

Five-fold cross validation was used to evaluate the methods performance, and training and testing samples were separated patient-wise. We report the average results over all folds in Table 1. For the sake of completion, we include the results achieved by related works. Unfortunately, to the best of our knowledge, there are no works reporting results for the setting we consider in our work for the INbreast dataset. The work in [HLCLT16] considers a similar setting, but for a database that is not originally digital. The very recent work in [DCB17], on the other hand, employs both MLO and CC images *jointly* for each patient. This implies having additional information, which we can consider leveraging in the future.

In the INbreast dataset, the naive baseline 1 algorithm obtained an AUC of 0.602, while the proposed architecture with the binary loss achieved an AUC of 0.630. This result is boosted to 0.650 once the AUC-maximization loss is employed. In the proprietary dataset, on the other hand, the results for the baseline 1 is of 0.677, while baseline 2 obtained an AUC of 0.727 and our complete method achieved an AUC of 0.767, representing a 9% and nearly 5% improvement, respectively. The advantage of the proposed formulation is also evident in the evolution of the AUC during training, shown in Figure 3 for one of the splits from the IMG dataset. The difference in performance between the two datasets can be attributed to the amount of training data in each of them: the proprietary data base contains about 2,000 images, while INbreast contains only 410, which might compromise the training of the model.

As can be seen from the results in Table 1, performing mammogram classification with only globally labeled data is significantly harder than having access to local annotations and masses segmentation. Lastly, while employing a different dataset (MIAS), it is worth noting that the work in [HK16] recently reported an AUC of 0.675 also employing a global label per image. Our method compares favorably with their results, in particular as their class criteria was selected as containing or not containing abnormalities. This might result in an easier task than discriminating malignant from benign abnormalities, as in our case.

6. Conclusion

In this work, we have presented a new deep learning architecture for the classification of globally-labeled mammographies. The particular design allows to circumvent the need of local annotations, as only a global label is used to train the entire model. The proposed CNN is trained by means of a new AUC-maximization loss as opposed to minimizing the classification error. Our approach is validated on two different datasets, showing that promising classification performance can be obtained in the very challenging scenario of globally-labeled mammograms.

As future work, a systematic analysis of the implications of the model used as the scanning network, and comparing different off-the-shelf trained models for this purpose, could potentially increase the overall performance. In addition, considering several input channels (for instance, with more than one mammography per patient, as done in [DCB17]) while leveraging the proposed AUC maximization loss, would likely yield improved results. On the other hand, one could employ the trained model presented in our work to improve other, more complex, algorithms such as those of localization. Finally, extending the experimental validation to other datasets will definitely contribute to the understanding of the capabilities and limitations of our methods. All these points are subjects of ongoing work.

References

[AS05] Ataman K., Street W. N.: Optimizing area under the roc curve using ranking SVMs. In *https://dollar.biz.uiowa.edu/∼street/research/kdd05kaan.pdf* (2005). 2

[BL08] Boyle P., Levin B.: *World cancer report 2008.* IARC Press, International Agency for Research on Cancer, 2008. 1

[CB12] Castro C. L., Braga A. P.: Improving ANNS performance on unbalanced data with an AUC-based learning algorithm. In *ICANN* (2012), Springer, pp. 314–321. 2

[CM03] Cortes C., Mohri M.: Auc optimization vs. error rate minimization. In *NIPS* (2003), vol. 9, p. 10. 1

[CSVZ14] Chatfield K., Simonyan K., Vedaldi A., Zisserman A.: Return of the devil in the details: Delving deep into convolutional nets. In *BMVC* (2014). 3

[DCB16] Dhungel N., Carneiro G., Bradley A. P.: The automated learning of deep features for breast mass classification from mammograms. In *MICCAI* (2016), Springer, pp. 106–114. 2, 4

[DCB17] Dhungel N., Carneiro G., Bradley A. P.: Fully automated classification of mammograms using deep residual neural networks. In *IEEE ISBI* (2017). 4, 5

[GJZZ13] Gao W., Jin R., Zhu S., Zhou Z.: One-pass auc optimization. In *ICML (3)* (2013), pp. 906–914. 2, 3

[HK16] Hwang S., Kim H.: Self-transfer learning for weakly supervised lesion localization. In *MICCAI* (2016), Springer, pp. 239–246. 2, 5

[HLCLT16] Huang C., Li Y., Change Loy C., Tang X.: Learning deep representation for imbalanced classification. In *IEEE CVPR* (2016), pp. 5375–5384. 1, 5

[HR04] Herschtal A., Raskutti B.: Optimising area under the roc curve using gradient descent. In *ICML* (2004), p. 49. 2

[JGWL16] Jiao Z., Gao X., Wang Y., Li J.: A deep feature based framework for breast masses classification. *Neurocomputing 197* (2016), 221–231. 2

[JY15] Jen C., Yu S.: Automatic detection of abnormal mammograms in mammographic images. *Expert. Syst. Appl. 42*, 6 (2015), 3048–3055. 2

[LCCM16] Li Y., Chen H., Cao L., Ma J.: A survey of computer-aided detection of breast cancer with mammography. *J. Health Med. Inf. 7(4)* (2016). 2

[LJ16] Lévy D., Jain A.: Breast mass classification from mammograms using deep convolutional neural networks. *arXiv preprint arXiv:1612.00542* (2016). 2

[MAD*12] 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. 4

[MW47] Mann H. B., Whitney D. R.: On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat* (1947), 50–60. 2

[PBP*08] Pal N. R., Bhowmick B., Patel S. K., Pal S., Das J.: A multi-stage neural network aided system for detection of microcalcifications in digitized mammograms. *Neurocomputing 71*, 13 (2008), 2625–2634. 2

[WSX16] Wang S., Sun S., Xu J.: Auc-maximized deep convolutional neural fields for protein sequence labeling. In *ECML PKDD* (2016), Springer, pp. 1–16. 2

[ZJYH11] Zhao P., Jin R., Yang T., Hoi S. C.: Online AUC maximization. In *(ICML)* (2011), pp. 233–240. 2, 3