DEEP LEARNING OF FEATURE REPRESENTATION WITH MULTIPLE INSTANCE LEARNING FOR MEDICAL IMAGE ANALYSIS

Yan Xu[1,2]*, Tao Mo*[2,3]*, Qiwei Feng*[2,4]*, Peilin Zhong*[2,4]*, Maode Lai*[5]*, Eric I-Chao Chang*[2*]

[1]State Key Laboratory of Software Development Environment,
Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education, Beihang University
[2]Microsoft Research, Beijing, China
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[4]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[5]Department of Pathology, School of Medicine, Zhejiang University, China
{eric.chang, v-tamo, v-qifen,v-pezhon}@microsoft.com, xuyan04@gmail.com, lmd@zju.edu.cn

ABSTRACT

This paper studies the effectiveness of accomplishing high-level tasks with a minimum of manual annotation and good feature representations for medical images. In medical image analysis, objects like cells are characterized by significant clinical features. Previously developed features like SIFT and HARR are unable to comprehensively represent such objects. Therefore, feature representation is especially important. In this paper, we study automatic extraction of feature representation through deep learning (DNN). Furthermore, detailed annotation of objects is often an ambiguous and challenging task. We use multiple instance learning (MIL) framework in classification training with deep learning features. Several interesting conclusions can be drawn from our work: (1) automatic feature learning outperforms manual feature; (2) the unsupervised approach can achieve performance that's close to fully supervised approach (93.56%) vs. (94.52%); and (3) the MIL performance of coarse label (96.30%) outweighs the supervised performance of fine label (95.40%) in supervised deep learning features.

***Index Terms*—** deep learning, feature learning, supervised, unsupervised, multiple instance learning

1. INTRODUCTION

In medical image analysis, it is common to design a group of specific features [1, 2] for a high-level task such as classification and segmentation [3]. Meanwhile, detailed annotation of medical images is often an ambiguous and challenging task. This paper addresses the effectiveness and efficiency of accomplishing high-level tasks with a minimum of manual annotation and good feature representations [4, 5, 6].

There is a rich body of literature on feature representation. The main methods of feature extraction are manually designed feature descriptors [7, 8], fully supervised feature learning [9] and unsupervised feature learning [10]. Manually designed feature descriptors [7, 11], including gradient operators and filter banks, are unable to capture complex variations frequently found in medical images. Fully supervised feature learning [9] requires a large amount of accurately annotated data. Obtaining such annotated data is time-consuming, labor-intensive, and ambiguous. Unsupervised feature learning [12, 13, 14, 15] is based on unlabeled data. It can learn intrinsic and subtle features from the statistics of the real data. In this paper, we study these methods in the medical image domain. We used SIFT [7], LBP [8] and L*a*b color histogram as manual features. We explored the features from the last hidden layer in deep learning neural networks as fully supervised features. We adopted the single-layer network of centroids derived from K-means clustering algorithm as unsupervised features [16]. Experiment results showed that both fully supervised and unsupervised feature learning are superior to manual features. In addition, we compared the influence of different numbers of nodes of the last hidden layer in fully supervised features. The high dimensional features are superior to the low dimensional features in the fully supervised feature learning.

In high-level tasks such as classification, weakly supervised methods combine the advantages of both the fully supervised and the unsupervised [3, 17]. The goal is to automatically extract fine-grained information from coarse-grained labels. Multiple Instance Learning is a particular form of weakly supervised method which we studied. A bag is comprised of many instances. Given a series of bag labels, MIL uses the bag labels (coarse-grained) to predict instance labels (fine-grained). In this paper, we study colon cancer classification based on histopathology images. A histopathology image is considered as a bag. An image is split into many patches as instances. A bag is labeled as positive if the bag contains at least one positive instance (cancer tissue). A bag is labeled as negative if the bag contains all negative instances.

This paper is organized as follows. In Section 2, we describe related work of feature learning and the MIL framework. In Section 3, we present the algorithms to study the efficiency and effectiveness of feature learning and weakly trained classifiers. In Section 4, we report experiment results from the different approaches. Then our conclusion is presented in Section 5.

2. RELATED WORK

Related work can be broadly divided into three categories: (1) medical image high-level tasks in medical imaging field, (2) deep learning in feature learning and classification, and (3) multiple instance

*Corresponding author. This work was supported by Microsoft Research (MSR). Thanks to MSR grant, Grant 61073077 from National Science Foundation of China, Grant SKLSDE-2011ZX-13 from State Key Laboratory of Software Development Environment in Beihang University, and the Fundamental Research Funds for the Central Universities of China.

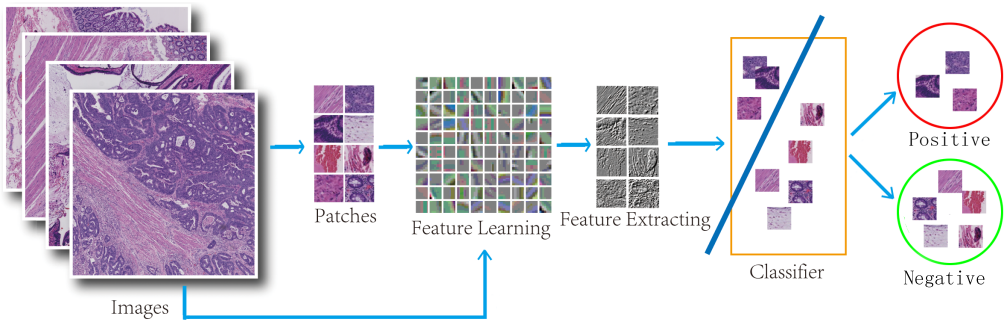

Fig. 1. The flow diagram of the algorithms with a minimum of manual annotation and good feature representations. The inputs include both cancer images and noncancer images. All images are used to generate patches. In feature learning processing, images/patches are used to downsample receptive fields. Feature learning is implemented by three methods containing full supervised deep learning, unsupervised learning of a single-layer network, and manual features. The next step is to extract features for each patches. In classifier processing, we conduct fully supervised classifier (SVM) and weakly supervised classifier (MIL). The overall patch-level classification (cancer vs. noncancer) can be obtained based on the confidences from classifiers. Red represents cancer patches while green represents noncancer patches.

learning.

High-level tasks such as classification and segmentation in medical imaging field is a hot topic. Due to clinical nature of the images, many previous work focus on feature design. The main methods consists of manually feature design, supervised feature learning, and unsupervised feature learning. Boucheron [18] and Chang [19] focus on manual feature design while Le [20] focuses on unsupervised feature learning. Boucheron *et al* [18] exploited segmentation results of cell nuclei as features to improve the classification accuracy in histopathology images of breast cancer. The feature dimension is 1035 in image-level classification. Chang *et al* [19] presented nuclear level morphometric features at various locations and scales within the spatial pyramid matching to classify tumor histopathology images. Le *et al* [20] proposed a two-layer network with nonlinear responses to automatically learn features in histopathology tumor images. In our work, we compared the three main methods on a colon histopathology dataset. Feature learning methods outperform manual feature operators.

Deep learning can be used for both classification and feature learning in various fields such as computer vision and speech. Deep learning as classifiers are used in acoustic emotion recognition [21] and object classes in ImageNet [22]. Deep learning can be used in feature learning including supervised [9] and unsupervised [20]. In our work, we attempted deep learning of feature representation with MIL to classify colon histopathology images.

Multiple Instance Learning is a weakly supervised learning framework. In training, the MIL framework utilizes a minimum of manual annotation. We previously proposed the framework to classify colon histopathology images [3, 17] by using the bag-level labeled data to predict the instance-level data. However, we used manual features with MIL to accomplish the task. In this paper, we combined deep learning of feature representation with the MIL framework to classify colon histopathology images. The algorithm combines training with minimal manual annotation and good feature representations. In addition, our method is general and can be applied to MIL tasks other than colon histopathology images.

3. ALGORITHMS

In this section, we describe the algorithms used in our experiments. Our task is to predict whether an image is positive (cancer) or negative (noncancer), and to outline cancer regions if it is positive. We formulate the problem as patch-level classification. If any patch in the image is recognized as positive, then the image will be considered as a cancer image. Otherwise, all patches belong to negative and the image is considered as a noncancer image. Our algorithm is a pipelined process as follows: (1) to produce patches from images of both positive (cancer) and negative (non-cancer), (2) to generate good feature representations using images/patches, (3) to extract features by learning feature models or manual feature operators, (4) to classify patches into positive or negative by using classifiers that are trained fully supervised or weakly supervised, and (5) to obtain the patch-level classification results. Fig. 1 is the algorithm diagram. We will introduce detailed descriptions of some key steps in the pipeline process.

3.1. Full supervised feature learning framework

In this section, we describe the algorithm for fully supervised deep learning of features. We propose a system based on deep learning having a set of linear filters in encoder and decoder. The network of deep learning is a process of deriving high-level features from low-level features. The nodes of low layers represent lower level features while the nodes of higher layers represent higher level features. The last hidden layer nodes can represent intrinsical features compared to lower layer features. Similar work can be found in [9] which was applied to speech recognition. We use the last hidden layer of deep learning as our fully supervised feature learning. Different networks can achieve different performances. Convolution and max/avg pool are considered as common layers of networks in image analysis.

In this paper, we attempt two networks for evaluating the efficiency and effectiveness of the last hidden layer features. In network one convolution and pool are alternately used without full connection layers (DNN2-F); in network two the last layer is the full connection one after convolution and pool (DNN1-F). The nodes generated by convolution and pool are enormous. In our experiment, the dimension number is 160,000. Principal Component Analysis (PCA) is used to reduce the dimension of the DNN features.

3.2. Unsupervised feature learning framework

Unsupervised feature learning is a method conducted without expensive manual annotation. It can learn intrinsic and subtle features

from the statistics of the real data [16]. Given the benefits of using unlabelled data, we explored unsupervised feature learning. In our experiment, we used the single-layer network of K-means centroids as the unsupervised feature learning. We describe feature learning and feature extraction respectively.

Feature learning: A receptive field (rf) is defined as a $d*d$ sub-image of a large $h*w$ image. The stride is set to 1 in our work, thus an image has $(h-d+1)*(w-d+1)$ receptive fields (rfs) totally. For a three-channel (RGB) image, a rf can be described as a vector in $\mathbb{R}^{3d^2}$. First step of the algorithm is to generate the "centroids" of the dataset. A centroid is also a vector in $\mathbb{R}^{3d^2}$ and the centroids are the "most common rfs" in all the images. We randomly extract n rfs from the image set and form P, and then run the K-means algorithm to generate k centroids $C_1, ..., C_k$. The K-means algorithm contains t iterations. In each iteration, we find the closest centroid for each rf in P, and assign the rf to the centroid. Then, for each centroid C_i, we take all rfs that assigned to this centroid in the current iteration, and modify the centroid into a new one C_i' which is the mean of all these rfs. After running such iteration for t rounds, the set of centroids converges to describe the most common rfs of P.

Feature extraction: The centroids are used to extract feature from an image. Suppose an image has a dimension of $h*w$, then it has $(h-d+1)*(w-d+1)$ rfs. For one rf $p \in \mathbb{R}^{3d^2}$, we can map it to a $\mathbb{R}^k$ vector $f(p)$, where

$$f_i(p) = \max\{0, \mu - z_i\}, \tag{1}$$

and $z_i = ||p - C^{(i)}||_2$, $\mu = (\sum_i z_i)/k$. Thus there are $(h-d+1)*(w-d+1)$ vectors in $\mathbb{R}^k$, then we divided the grids into four equal parts, and sum up the vectors in each part to obtain 4 vectors in $\mathbb{R}^k$, which can be concentrated into a $4k$-dimension vector. This is the feature of the input image.

Notice that we do not use any label information in the K-means algorithm and the feature extracting process.

3.3. Multiple Instance Learning

Detailed manual annotations are time-consuming and intrinsically ambiguous. An alternative is to learn local concepts using global annotations, which is the main idea of Multiple Instance Learning (MIL). MIL is a weakly supervised learning framework. The training set contains labeled bags that are composed of unlabeled instances, and the task is to predict the labels of unseen bags and instances. In this paper, a bag is a large size image and an instance is a distinguishable patch. The bag is labeled positive if and only if there is at least one positive instance in the bag, i.e. some part of the image, but maybe not the whole image are cancer tissue. Thus we can formulate a binary MIL model which optimizes the loss function of bag classification, while the bag classifier is a softmax of the instance classifiers. Specifically, $X_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ is the i^{th} bag in the training set, m is the number of instances in the i^{th} bag and $\{x_{i1}, x_{i2}, \ldots, x_{im}\}$ are instances of this bag. $y_i \in \{-1, +1\}$ is the label, -1 means negative bag and +1 means positive bag. $H(X) \in X \to [0,1]$ and $h(x) \in x \to [0,1]$ are bag-level classifiers and instance-level classifiers, which give the positive probability of bags and instances. The loss function is:

$$\mathcal{L}(H) = -\sum_{i=1}^{n} \mathbf{1}(y_i = 1) \log H(X_i) + \mathbf{1}(y_i = -1) \log\left(1 - H(X_i)\right), \tag{2}$$

where $\mathbf{1}(\cdot)$ is an indicator function.

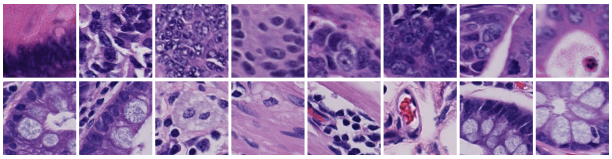

Fig. 2. Several examples from fully supervised dataset. the top row: positive (cancer); the bottom row: negative (noncancer).

Using the gradient descent algorithm, we can iteratively train weak classifiers $h'(x)$ using weights:

$$w_{ij} = -\frac{\partial \mathcal{L}(H)}{\partial h(x_{ij})} = -\frac{\partial \mathcal{L}(H)}{\partial H(X_i)} \frac{\partial H(X_i)}{\partial h(x_{ij})}, \tag{3}$$

and updates $h(x)$ by $h(x) \leftarrow h(x) + \alpha h'(x)$, where α is the coefficient which is obtained by line searching to minimize the loss function. After sufficient iterations for the loss function to converge, we generate an efficient classifier. This algorithm is called MIL-Boost.

4. EXPERIMENTS

4.1. Datasets

High resolution histopathology images are used to construct our datasets. All the images were selected from histopathology images of 132 patients. Each image is set to be 10000x10000 pixels due to the computing power of a single machine. This is a bag mentioned in MIL. We sampled 200x200 pixels patches while overlap step size is 100 pixels, thus we obtained 9801 patches in an image, each patch is an instance. Detailed datasets are as follows (See Table 1):

Fully supervised dataset: First we chose 30 images with cancer and labeled cancer regions, then 9000 patches completely enclosed within the labeled cancer regions were used as positive instances. From 30 noncancer images, we randomly sampled 9000 patches as negative instances. The *Training Set* and *Testing Set*, both containing 4500 positive instances and 4500 negative instances were randomly selected from the above data. The *Training Set* was not only used to train fully supervised learning algorithm like SVM and DNN, but was also the evaluation dataset for weakly supervised learning. Fig. 2 shows several example of patches.

Weakly supervised dataset: 30 positive images and 83 negative images were used as the *Bags Set*, each image contained 9801 patches, thus we had 113 labeled bags and over 1 million unlabeled instances to build the MIL model.

Annotations: Both fully supervised annotation (cancer region) and weakly-supervised annotation (the label of bags) were labeled by two pathologists independently. When there was a disagreement, a third senior pathologist would discuss with them to determine the ground truth.

Table 1. Datasets Distribution

Dataset	Positive		Negative	
	Instances	Bags	Instances	Bags
Training Set	4,500	N/A	4,500	N/A
Testing Set	4,500	N/A	4,500	N/A
Bags Set	294,030	30	813,483	83

4.2. Settings

We studied four different types of features on all 200x200 patches, and finished classification of fully supervised learning and of weakly supervised learning on them.

Feature extraction: Following methods were used

Manual Feature (MF): Generic object classification features were chosen, including SIFT, LBP, and L*a*b color histogram. Feature dimension is 188.

K-means: To achieve good representation of the feature space, we randomly sampled 10 million 8x8 receptive fields from *Bags Set*, then we clustered them into 1600 centroids and obtained 4x1600=6400 dimension features for each instance.

DNN1-F: We trained the network 3x200x200-32C5-MP2-32C5-MP2-64C5-MP2-1000N-2N on *Training Set*, and applied the optimized network on each patch to obtain features. The last full connection layer is used to extract features, and feature dimension is 1000.

DNN2-F: Similar with DNN1-F, except using a different network 3x200x200-32C5-MP2-32C5-MP2-64C5-MP2-2N. The last conv3 layer is used to extract features, feature dimension is $160,000$. Due to the enormous dimension length, PCA was performed to compress the dimension of the DNN features to 1000 dimensions, and following experiments were carried out on reduced feature.

DNN1-C: The same features with DNN1-F.

DNN2-C: The same features with DNN2-F.

Fully supervised learning: Linear SVM with default parameters were trained on *Training Set*. The classifier was used in MF, K-means, DNN1-F, and DNN2-F. We also presented the DNN classification results of the two neural networks above (DNN1-C and DNN2-C), using the same training set.

Weakly supervised learning: MIL-Boost algorithm was used for weakly supervised learning, the softmax function was Generalized Mean (GM) with $r = 5$, weak classifiers were Decision Stump and Decision Interval, we ran 5000 iterations or until the loss function converged. *Bags Set* were used for training and the model was tested on *Training Set* to find the best threshold for the testing process.

There are 1,107,513 patches in all, the dimensions and data sizes of these features are presented in Table 2. K-means feature takes storage far larger than DNN1-F feature, because the latter has a smaller dimension and is sparser than the former.

Table 2. Dimensions and data sizes for different features (GB)

	Dimension	Data Size
MF	188	2.85
K-means	6400	123.72
DNN1-F	1000	6.70
DNN2-F	160000	265.52
DNN2-F (after PCA)	1000	14.71

4.3. Results

The accuracies on *Testing Set* of all experiments above are presented in Table 3. DNN2-F benefitted from the detailed representation of high dimension feature and showed the best accuracy. Weakly supervised learning on K-means feature was the most interesting part, both feature extraction and training phase didn't require instance labels but it performed better than the manual features. With even more unlabeled data, this approach may result in classification performance approaching that of fully supervised training approach.

In fully supervised classification, the performances of DNN-Fs were similar with DNN-Cs. Among these methods, K-means feature, which is simple and has few parameters, approached the DNN1-F in accuracy. It gives support to unsupervised feature extracting.

Table 3. The Performances of various competing algorithms

	Full Supervised	Weakly Supervised
MF	91.52%	87.28%
K-means	93.56%	89.43%
DNN1-F	94.52%	96.30%
DNN2-F	97.81%	97.44%
DNN1-C	95.40%	N/A
DNN2-C	97.30%	N/A

4.4. Comparing Different Features

To finish the experiments in reasonable amount of time, both feature extraction and model learning were implemented with Message Passing Interface (MPI) and carried out on Windows High Performance Computing (HPC) Cluster. We used up to 128 compute nodes each with 8 processors and 16 GB of RAM. For DNN training and feature extraction, we used 4 servers each with 24 processers, 72 GB of RAM and 2 NVIDIA Tesla M2090 GPU cards.

Time cost of each phase for four feature sets can be seen in Table 4. Preprocess for K-means feature is clustering and Preprocess for DNN feature is training neural network. Feature extraction for manual feature and K-means feature were distributed and the value was the time that one compute node needed to handle one piece of 10000x10000 image. The framework of MIL-Boost is well parallelized.

Among these feature extraction methods, manual feature is the fastest but least accurate and it must be well designed for different datasets. K-means feature is totally unsupervised in the extraction phase and represents the dataset in a robust and efficient way. However, high computation complexity and high feature dimension do not fit well with large scale data. DNN feature is the most accurate one but must be trained with fully labeled data.

Table 4. Time costs for different features (hours)

	Preprocess	Feature Extracting	MIL-Boost
MF	N/A	0.02	0.88
K-means	2	5	6.3
DNN1-F	4.6	0.17	1.6
DNN2-F	4.4	0.22	1.7

5. CONCLUSION

In this paper, we propose an algorithm with a minimum of manual annotation and good feature representations to accomplish high-level tasks such as classification and segmentation in medical image analysis. We compared four experiments of feature representations on the dataset consisting of colon cancer histopathology images. The experiment results demonstrated that feature learning is superior to manual feature operators. The performance of unsupervised feature learning (93.56%) approaches the performance of fully supervised feature learning (94.52%) in the fully supervised classification.

Furthermore, the MIL framework is effective and efficient in classification. In supervised deep learning features, the MIL performance of coarse label (96.30%) exceeds the supervised performance of fine label (95.40%).

Due to features generated by the limited amount of unlabeled data and the single-layer network in the unsupervised feature learning, unsupervised feature performance is slightly worse than supervised. For future work, we will conduct an experiment with more unlabeled data and the multi-layer network in unsupervised feature learning. In addition, we will explore using an auto-encoding DNN instead of K-means for learning feature representation without fully labeled data.

6. REFERENCES

[1] P.W. Huang and C.H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *TMI*, vol. 28, no. 7, pp. 1037–1050, 2009.

[2] J. Kong, O. Sertel, H. Shimada, K.L. Boyer, J.H. Saltz, and M.N. Gurcan, "Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation," *Pattern Recognition*, vol. 42, no. 6, pp. 1080–1092, 2009.

[3] Y. Xu, J.Y. Zhu, E. Chang, and Z. Tu, "Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering," in *CVPR*, 2012, pp. 964–971.

[4] G.E. Hinton, S. Osindero, and Y.W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[5] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *ICCV*, 2009, pp. 2146–2153.

[6] H. Lee, C. Ekanadham, and A. Ng, "Sparse deep belief net model for visual area v2," in *NIPS*, 2007, pp. 873–880.

[7] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[8] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *TPAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[9] Z.J. Yan, Q. Huo, and J. Xu, "A scalable approach to using dnn-derived features in gmm-hmm based acoustic modeling for lvcsr," in *ISCA*, 2013.

[10] M. Ranzato, Y. Boureau, and Y. LeCun, "Sparse feature learning for deep belief networks," in *NIPS*, 2007, pp. 1185–1192.

[11] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006, vol. 2, pp. 2169–2178.

[12] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *CVPR*, 2008, pp. 1–8.

[13] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun, "Learning invariant features through topographic filter maps," in *CVPR*, 2009, pp. 1605–1612.

[14] H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *ICML*, 2009, pp. 609–616.

[15] Y. Bengio, A.C. Courville, and P. Vincent, "Unsupervised feature learning and deep learning: A review and new perspectives," *CoRR*, 2012.

[16] A. Coates, A.Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *ICAIS*, 2011, pp. 215–223.

[17] Y. Xu, J.W. Zhang, E. Chang, M.D. Lai, and Z.W. Tu, "Contexts-constrained multiple instance learning for histopathology image analysis," in *MICCAI*, 2012, pp. 623–630.

[18] L.E. Boucheron, B. Manjunath, and N.R. Harvey, "Use of imperfectly segmented nuclei in the classification of histopathology images of breast cancer," in *ICASSP*, 2010, pp. 666–669.

[19] H. Chang, A. Borowsky, P. Spellman, and B. Parvin, "Classification of tumor histology via morphometric context," in *CVPR*, 2013, pp. 2203–2210.

[20] Q.V. Le, J. Han, J.W. Gray, P.T. Spellman, A. Borowsky, and B. Parvin, "Learning invariant features of tumor signatures," in *ISBI*, 2012, pp. 302–305.

[21] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *ICASSP*, 2011, pp. 5688–5691.

[22] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1106–1114.