# Cancer Metastasis Detection With Neural Conditional Random Field

**Yi Li**
Baidu Silicon Valley Artificial Intelligence Lab
1195 Bordeaux Dr. Sunnyvale, CA 94089
liyi17@baidu.com

**Wei Ping**
Baidu Silicon Valley Artificial Intelligence Lab
1195 Bordeaux Dr. Sunnyvale, CA 94089
pingwei01@baidu.com

## Abstract

Breast cancer diagnosis often requires accurate detection of metastasis in lymph nodes through Whole-slide Images (WSIs). Recent advances in deep convolutional neural networks (CNNs) have shown significant successes in medical image analysis and particularly in computational histopathology. Because of the outrageous large size of WSIs, most of the methods divide one slide into lots of small image patches and perform classification on each patch independently. However, neighboring patches often share spatial correlations, and ignoring these spatial correlations may result in inconsistent predictions. In this paper, we propose a neural conditional random field (NCRF) deep learning framework to detect cancer metastasis in WSIs. NCRF considers the spatial correlations between neighboring patches through a fully connected CRF which is directly incorporated on top of a CNN feature extractor. The whole deep network can be trained end-to-end with standard back-propagation algorithm with minor computational overhead from the CRF component. The CNN feature extractor can also benefit from considering spatial correlations via the CRF component. Compared to the baseline method without considering spatial correlations, we show that the proposed NCRF framework obtains probability maps of patch predictions with better visual quality. We also demonstrate that our method outperforms the baseline in cancer metastasis detection on the Camelyon16 dataset.

## 1 Introduction

Breast cancer is one of the leading causes of death among women in the United States [22]. Early cancer diagnosis and treatment play a crucial role in improving patients' survival rate [20]. One of the most important early diagnosis is to detect metastasis in lymph nodes through microscopic examination of hematoxylin and eosin (H&E) stained histopathology slides. In recent years, pathologists have been using Whole-slide Images (WSIs) to distinguish between normal and tumor cells and localize malignant lesions [4]. However, manually detecting tumor cells within extremely large WSIs (e.g., 100,000 $\times$ 200,000 pixels) can be tedious and time-consuming. Significant discordance on detection results among different pathologists has also been reported [7]. Therefore, various computer-aided diagnosis (CAD) systems have been developed to assist pathologists to detect cancer metastasis in WSIs [6, 3].

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.

In recent years, deep convolutional neural networks (CNNs) have shown significant improvements on a wide range of computer vision tasks on natural images, e.g. image classification [16, 21, 11], object detection [10, 9], and semantic segmentation [18]. Similarly, a few promising studies have also applied deep CNNs to analyse medical images and particularly WSIs [24, 8, 12, 17, 23, 28, 27], among which [24] won the Camelyon16 challenge [1] for metastasis detection. Because of the extremely large size of WSIs, most of the studies first extracted small patches (e.g. $256 \times 256$ pixels) from WSIs, and trained a deep CNN to classify these small patches into normal or tumor regions. A probability map of the original WSI being tumor or normal at patch level was later obtained and metastasis detection was performed based on this probability map. However, the small patches and their neighbors often share spatial correlations. Because the patches were extracted and trained independently, the spatial correlations were not modeled explicitly. Therefore, during inference time, the predictions over neighboring patches may be inconsistent, and the patch level probability map may contain isolated outliers [14, 25].

To explicitly model the spatial correlations between neighboring patches, Kong et al. [14] recently proposed Spatio-Net that uses 2D Long Short-Term Memory (LSTM) layers to capture the spatial correlations based on patch features extracted from a CNN classifier. However, Spatio-Net uses a two-stage training approach, and therefore the CNN feature extractor is not aware of the spatial correlations [25]. In parallel to our work and very recently, we notice a similar work [25] that also uses features extracted from a CNN classifier to represent neighboring patches. Conditional random field (CRF) is then applied on these patch features to model spatial correlations and refine the predicted probability map during a post-processing stage. In addition to the same issue due to the two stage framework, there is a significant computational overhead during the CRF post-processing, and the authors have to select a limited number of features (e.g., 5 reported in [25]) from the original high dimensional patch representations to perform the CRF inference algorithm on CPU.

In this paper, we propose an alternative method for modeling spatial correlations between neighboring patches through neural conditional random field (NCRF). NCRF is a probabilistic graphical model based on the idea of conditional random fields as recurrent neural networks [26] that directly incorporates a fully connected CRF on top of the CNN feature extractor. The marginal label distribution of each patch is obtained through the mean-field approximate inference algorithm. The whole deep network can be trained in an end-to-end manner with the standard back-propagation algorithm, avoiding the post-processing stage. Because the mean-field inference algorithm is also performed on GPU, the CRF component introduces minor computational overhead and allows very large feature dimensions, e.g. 4096 from the VGG-19 architecture [21]. The CNN feature extractor also benefits from jointly training with the CRF component, since it is now aware of the spatial correlations between neighboring patches. Compared to the baseline method that does not consider patch spatial correlations, we show that, 1) NCRF improves the visual quality of the probability map, 2) NCRF improves the CNN feature extractor, and 3) NCRF improves the performance of cancer metastasis detection.

## 2 Method

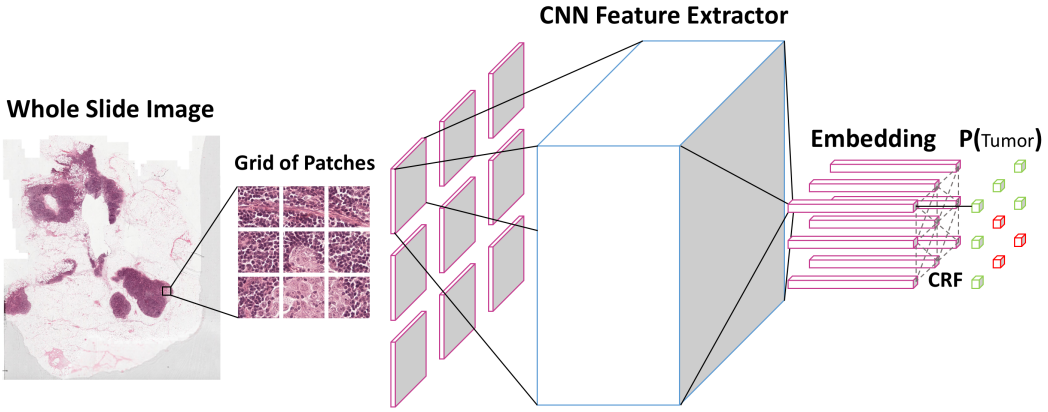


Figure 1: The architecture of NCRF model.

In this section, we describe the details of the proposed neural conditional random field (NCRF) model. Figure 1 shows the overall architecture of NCRF. It has two major components: CNN and CRF. The CNN component acts as a feature extractor, that takes a grid of patches as input, and encodes each patch as a fixed-length vector representation (i.e. embedding). The CRF component takes the grid of embeddings as input and models their spatial correlations. The final output from the CRF component is the marginal distribution of each patch being normal or tumor given the grid of patch embeddings. We illustrate the details of each component in the next two sections.

### 2.1 Patch Embedding With CNN

To extract comprehensive feature representation of each patch, we employ two classical deep CNN architectures that have proven to be powerful in image classification task, AlexNet [16] and VGG-19 [21]. We slightly modified the original VGG-19 architecture by replacing the first two 3×3 stride 1 convolution layers with one 7×7 stride 2 convolution layer. For each architecture, we use the activations from the second last fully connected layer before the softmax layer as the embedding for each patch. The embedding size is 4096 for both AlexNet and VGG-19, which is much larger than the embedding size of 5 reported in [25] after feature selection.

### 2.2 Spatial Modeling With CRF

In this section, we describe the methodology details of the CRF component. We denote a grid of patch embeddings obtained from CNN as $\boldsymbol{x} = \{x_i\}_{i=1}^N$, where $N$ is the number of patches within the grid, e.g. 25 for a grid of $5 \times 5$. Let $\mathbf{Y} = \{y_i\}_{i=1}^N$ be the random variables associated with each patch $i$, that represents the label of patch $i$ takes a value from $\{normal,\ tumor\}$. The conditional distribution $P(\mathbf{Y} \mid \boldsymbol{x})$ can be modeled as a CRF with a Gibbs distribution of

$$P(\mathbf{Y} = \boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z(\mathbf{x})} \exp(-E(\boldsymbol{y}, \boldsymbol{x})) \tag{1}$$

where $E(\boldsymbol{y}, \boldsymbol{x})$ is the energy function that measures the cost of $\mathbf{Y}$ taking a specific configuration $\boldsymbol{y}$ given $\boldsymbol{x}$, and $Z(\boldsymbol{x})$ is the partition function that insures $P(\mathbf{Y} = \boldsymbol{y} \mid \boldsymbol{x})$ is a valid probability distribution. In a fully-connected pairwise CRF [15], the energy function is given by:

$$E(\mathbf{y}, \boldsymbol{x}) = \sum_i \psi_u(y_i) + \sum_{i<j} \psi_p(y_i, y_j) \tag{2}$$

where $i, j$ ranges from 1 to $N$. $\psi_u(y_i)$ is the unary potential that measures the cost of patch $i$ taking the label $y_i$ given the patch embedding $x_i$, and $\psi_p(y_i, y_j)$ is the pairwise potential that measures the cost of jointly assigning patch $i, j$ with label $y_i, y_j$ given the patch embeddings $x_i, x_j$. Pairwise potential $\psi_p(y_i, y_j)$ models spatial correlations between neighboring patches, and would encourage low cost for assigning $y_i, y_j$ with the same label if $x_i, x_j$ are similar. We implement the unary potential $\psi_u(y_i)$ as the negative log-likelihood of patch $i$ taking label $y_i$, which is the negative logit for label $y_i$ before the softmax layer of the CNN classifier. We implement the pairwise potential as the weighted cosine distance between $x_i, x_j$:

$$\psi_p(y_i, y_j) = \mathbb{I}(y_i = y_j) \cdot w_{i,j} \left(1 - \frac{x_i \cdot x_j}{\|x_i\| \, \|x_j\|}\right) \tag{3}$$

where $\mathbb{I}(y_i = y_j)$ is the indicator function that checks the label compatibility between $y_i, y_j$, and $w_{i,j}$ is a trainable weight which controls the correlation strength between two patches $i, j$ within the grid. Typically, fully connected CRF also includes another distance term for pairwise potential that encodes the spatial distance between two patches $i, j$ [15]. However we did not observe clear improvements by including such distance term, and if we put a trainable coefficient before the term, the coefficient was pushed to zero during training. On the other hand, we observed the trainable weight $w_{i,j}$ correlated well with the relative distances between different patches after model converged.

In order to train the CNN-CRF architecture end-to-end with the standard back-propagation algorithm, we need to obtain the marginal distribution of each patch label $y_i$, so that it can be used to compute the cross-entropy loss with respect to the ground truth labels [26]. However, exact marginal inference is intractable, and we use mean-field approximate inference, where the original CRF distribution $P(\mathbf{Y})$[1]

[1] We omit the dependency on $\boldsymbol{x}$ here for clarity.

is approximated with a simpler distribution $Q(\mathbf{Y})$, that can be written as the product of marginal distributions of each individual patch $i$, $Q(\mathbf{Y}) = \prod_i^N Q_i(y_i)$. By minimizing the KL divergence between $Q(\mathbf{Y})$ and $P(\mathbf{Y})$, $\mathbb{KL}(Q(\mathbf{Y}) \| P(\mathbf{Y}))$, we derive the update step for each marginal distribution $Q_i(y_i)$ [19]:

$$\log Q_i(y_i) = \mathbb{E}_{-Q_i}\left[\log \tilde{P}(\mathbf{Y})\right] + \text{const} \tag{4}$$

where $\mathbb{E}_{-Q_i}[f(\mathbf{Y})]$ means taking the expectation of $f(\mathbf{Y})$ with respect to all the variables except $y_i$, and $\tilde{P}(\mathbf{Y}) = \exp(-E(\boldsymbol{y}, \boldsymbol{x}))$ is the unnormalized CRF distribution. The mean-field inference algorithm is summarized in Algorithm 1

**Algorithm 1** Mean-field inference algorithm

compute $\psi_u(y_i)$ for all $i$ and $\psi_p(y_i, y_j)$ for all $i, j$
$\log \tilde{P}(\mathbf{Y}) \leftarrow -\left[\sum_i \psi_u(y_i) + \sum_{i<j} \psi_p(y_i, y_j)\right]$
initialize $Q_i(y_i) \leftarrow \exp(-\psi_u(y_i))$ for all $i$
normalize $Q_i(y_i)$ for all $i$
**for** T iterations **do**
  $\log Q_i(y_i) \leftarrow \mathbb{E}_{-Q_i}\left[\log \tilde{P}(\mathbf{Y})\right]$ for all $i$
  normalize $Q_i(y_i)$ for all $i$
**end for**

Finally, after a fixed number of mean-field iterations, we use the approximate marginal distribution of each patch label $Q_i(y_i)$ to compute the cross-entropy loss and train the whole model with back-propagation algorithm.

## 3 Experiments

In this section, we present empirical evaluations on the proposed NCRF method. We demonstrate its advantages over the baseline method without CRF in three aspects: 1) NCRF obtains smoother probability maps with sharp boundaries than the baseline method, 2) NCRF achieves higher patch level classification accuracies from the CNN feature extractor than the baseline method, and 3) NCRF outperforms the baseline method in cancer metastasis detection.

### 3.1 Data Preparation

We conducted all the experiments based on the Camelyon16 dataset[1, 5]. This dataset includes 160 normal and 110 tumor WSIs for training, 81 normal and 49 tumor WSIs for testing. Slides were exhaustively annotated by pathologists in pixel level, with a few exceptions reported in [17]. We conducted all the experiments on $40\times$ magnification. To exclude the background regions of each training slide, we split each slide into patches of $256 \times 256$ pixels and exclude patches where the mean of gray values is greater than 204 (0.8 of 255) and the variance of gray values is less than 100 [13]. We used Normal_001 to Normal_140 and Tumor_001 to Tumor_100 for training, and the rest of training slides for validation. To generate patches, we first randomly picked one slide and then randomly sampled a coordinate from the slide as the center of the patch. We randomly sampled $100,000$ $768 \times 768$ patches from the tumor regions of the tumor slides as positive samples. We randomly sampled $50,000$ $768 \times 768$ patches from the non-tumor non-background regions of the tumor slides and $50,000$ $768 \times 768$ patches from the non-background regions of the normal slides as negative samples. This is equivalent to have $0.9$ millions $256 \times 256$ patches for both normal and tumor regions. We also sampled $10,000$ $768 \times 768$ patches from both normal and tumor slides for validation using the same strategy.

### 3.2 Implementation Details

NCRF was implemented with TensorFlow [2] and trained with NVIDIA GeForce GTX 1080 Ti GPU. Note that, the CRF component introduces less than 0.1 seconds of computation overhead per batch iteration, since the mean-field inference algorithm is also performed on GPU. During the training

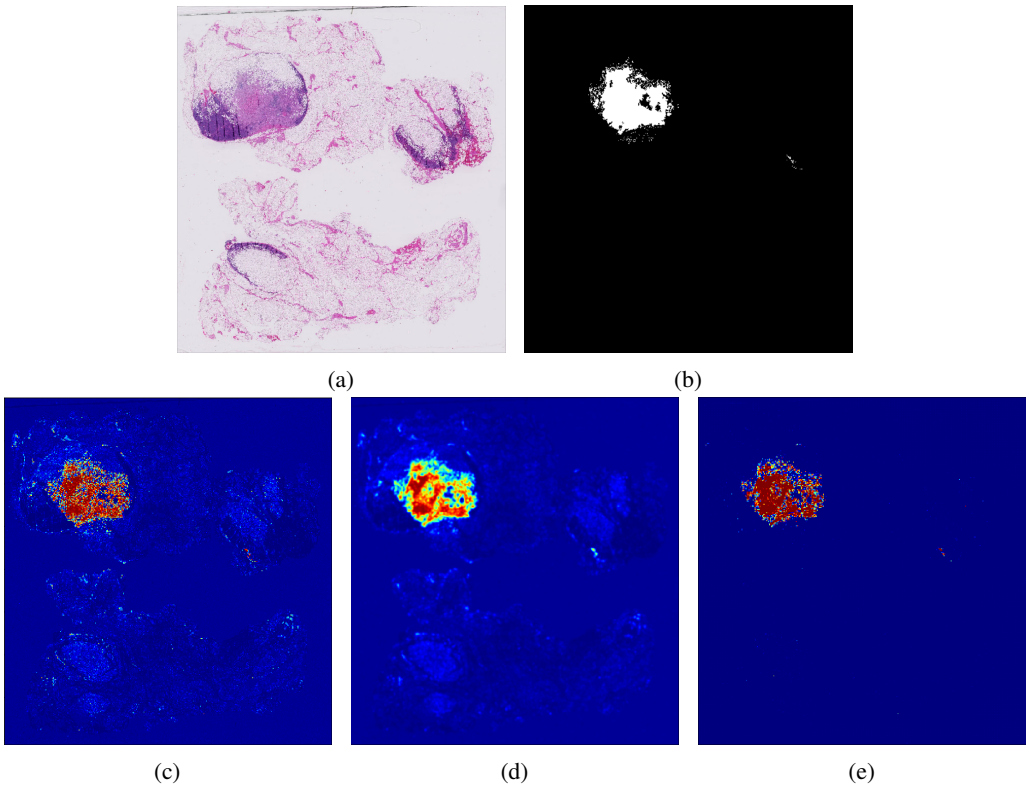


Figure 2: Predicted probability maps of Test_026 from AlexNet. (a) original WSI, (b) ground truth annotation, (c) baseline method, (d) baseline method with Gaussian smoothing, (e) NCRF

stage, a batch size of 20 $768 \times 768$ patches were feed into the model. Each $768 \times 768$ patch was further split into a $3 \times 3$ grid of $256 \times 256$ patches and their corresponding labels were retrieved. The pixel values of patches were normalized by subtracting the mean and dividing the standard deviation of each channel. We used stochastic gradient descent of learning rate $0.001$ and a momentum of $0.9$ to optimize all the architectures. A dropout rate of $0.5$ was applied to fully connected layers of all the architectures to prevent over-fitting. The mean-field inference algorithm was performed 10 iterations for all the architectures.

### 3.3 NCRF Obtains Smooth Probability Maps

Figure 2 shows the predicted probability maps of Test_026 from the baseline method, baseline method with Gaussian smoothing, and NCRF, all based on the AlexNet architecture. We can see the probability map from the baseline method that does not consider spatial correlations tends to contain isolated outlier predictions. Simple smoothing, e.g. Gaussian filtering, that takes the weighted average of predictions from surrounding patches [14], can effectively eliminate outlier predictions. However, it also tends to blur boundaries of tumor regions and may eliminate predictions of micrometastasis, e.g. the much smaller region on the right hand side of the ground truth annotation in Figure 2b. Compared to the baseline method and simple Gaussian smoothing, NCRF obtains a much smoother probability map while keeps sharp boundaries. The small micrometastasis prediction is also retained. From a practical point of view for computer-aided diagnosis systems, smoother probability maps allow pathologists to focus more on potential tumor regions and less on isolated outliers.

### 3.4 NCRF Improves CNN Feature Extractor

NCRF improves the CNN feature extractor by incorporating spatial correlations between neighboring patches during training. Table 1 shows the patch classification accuracies of the baseline method and NCRF on the validation set. NCRF consistently improves the patch classification accuracies on both AlexNet and VGG-19, compared to the baseline method without considering spatial correlations.

Table 1: Patch classification accuracies on the validation set.

| | baseline | NCRF |
|---|---|---|
| AlexNet[16] | 0.907 | 0.928 |
| VGG-19[21] | 0.911 | 0.929 |

Table 2: Average FROC score on the test set.

| | baseline | NCRF |
|---|---|---|
| AlexNet[16] | 0.511 | 0.538 |
| VGG-19[21] | 0.563 | 0.576 |

This result shows the CNN feature extractor benefits from the end-to-end joint training with CRF, compared to the previous two-stage training framework proposed in [14, 25], where the CNN feature extractor is not aware of the patch spatial correlations.

### 3.5 NCRF Improves Cancer Metastasis Detection

We evaluated the performance of cancer metastasis detection of NCRF based on the average free-response receiver operating characteristic (FROC) score on the Camelyon16 test set. Given a list of predicted coordinates of cancer metastasis, the average FROC score is defined as the average detection sensitivity at 6 predefined false-positive rates per slide: $1/4$, $1/2$, $1$, $2$, $4$ and $8$. Higher average FROC score means better detection performance. We used the same strategy in [17], to obtain the coordinates of cancer metastasis based on a given probability map.

Table 2 shows the average FROC scores of the baseline method and NCRF on the test set based on AlexNet and VGG-19. NCRF also consistently improves the average FROC score on both AlexNet and VGG-19, compared to the baseline method.

We were not able to reproduce the state-of-art baseline average FROC score of $0.705$ reported in [24], as no open source codebase was released from the original authors. One of the main reasons may be that we could not achieve significantly better patch classification accuracy using VGG-19 than AlexNet, as reported in [24].

## 4 Discussion

In this paper, we propose a neural conditional random field (NCRF) framework to detect cancer metastasis in Whole-slide Images (WSIs). NCRF is able to consider the spatial correlations between neighboring patches through the fully connected CRF component. Compared to previous methods, the CRF component is unified with the CNN feature extractor, and the whole model can be trained end-to-end with standard back-propagation algorithm. Because of this joint training framework, the CNN feature extractor also benefits from considering the spatial correlations while the CRF component introduces minor computational overhead. Compared to the baseline method without considering patch spatial correlations, NCRF obtains not only smoother probability maps but also better performances in cancer metastasis detection. We note our current baseline average FROC does not achieve the state-of-art result, and one of our future work is focusing on improving out baseline performance. Another future direction is using a grid of more than $3 \times 3$ patches as input, since it corresponds to a larger receptive field and may achieve better performance in cancer metastasis detection.

## References


[1] Camelyon 2016. 2016. https://camelyon16.grand-challenge.org.

[2] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning.

[3] G. Apou, B. Naegel, G. Forestier, F. Feuerhake, and C. Wemmert. Efficient region-based classification for whole slide images. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics*, pages 239–256. Springer, 2014.

[4] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho. Classification of breast cancer histology images using convolutional neural networks. *PloS one*, 12(6):e0177544, 2017.

[5] B. E. Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. van der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[6] R. F. Brem, J. A. Rapelyea, G. Zisman, J. W. Hoffmeister, and M. P. DeSimio. Evaluation of breast cancer with a computer-aided detection system by mammographic appearance and histopathology. *Cancer*, 104(5):931–935, 2005.

[7] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, 313(11):1122–1132, 2015.

[8] B. Geçer. *Detection and classification of breast cancer in whole slide histopathology images using deep convolutional networks*. PhD thesis, Bilkent University, 2016.

[9] R. Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2424–2433, 2016.

[13] A. Janowczyk and A. Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.

[14] B. Kong, X. Wang, Z. Li, Q. Song, and S. Zhang. Cancer metastasis detection via spatially structured deep network. In *International Conference on Information Processing in Medical Imaging*, pages 236–248. Springer, 2017.

[15] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*, 2017.

[18] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[19] K. P. Murphy. Machine learning: a probabilistic perspective. 2012.

[20] S. Saadatmand, R. Bretveld, S. Siesling, and M. M. Tilanus-Linthorst. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. *Bmj*, 351:h4901, 2015.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] U.S. Cancer Statistics Working Group. United states cancer statistics: 1999-2014 incidence and mortality web-based report. *Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute*, 2017. Available at: www.cdc.gov/uscs.

[23] Y. S. Vang, Z. Chen, and X. Xie. Deep learning framework for multi-class breast cancer histology image classification. *arXiv preprint arXiv:1802.00931*, 2018.

[24] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[25] F. G. Zanjani, S. Zinger, et al. Cancer detection in histopathology whole-slide images using conditional random fields on deep embedded spaces. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 105810I. International Society for Optics and Photonics, 2018.

[26] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.

[27] W. Zhu et al. Adversarial deep structured nets for mass segmentation from mammograms. In *IEEE ISBI*, 2018.

[28] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–611. Springer, 2017.