

# Manifold learning with graph-based features for identifying extent of lymphocytic infiltration from high grade, HER2+ breast cancer histology

Ajay Basavanhally<sup>1</sup>, Shannon Agner<sup>1</sup>, Gabriela Alexe<sup>2</sup>, Gyan Bhanot<sup>3</sup>, Shridar Ganesan<sup>3</sup>, and Anant Madabhushi<sup>1</sup>

<sup>1</sup>Department of Biomedical Engineering, Rutgers University, Piscataway, NJ, USA

<sup>2</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA

<sup>3</sup>The Cancer Institute of New Jersey, New Brunswick, NJ, USA

**Abstract**—It has been proposed that molecular changes in breast cancer (BC) may be accompanied by corresponding changes in phenotype. One such phenotype is the presence of lymphocytic infiltration (LI), a form of immune response seen often in high grade BC. The presence of LI in BC histology has been shown to correlate with prognosis and course of treatment. The advent of digitized histopathology has made tissue slides amenable to computer aided diagnosis (CAD). While texture-based features have recently been shown to successfully distinguish between tissue classes in histopathology, the similarity in appearance of BC nuclei and LI suggests that texture features alone may be insufficient. In this paper, we present a methodology that integrates manifold learning with graph-based features to distinguish high grade BC histology specimens based on the presence or absence of LI. Lymphocytes are first automatically detected via a segmentation scheme comprising a Bayesian classifier and template matching. For a total of 41 samples, the graph-based features, in conjunction with a Support Vector Machine classifier, achieve a classification accuracy of 89.50%. Our method is also compared against the popular Varma-Zisserman (VZ) texon-based classifier, which achieves a maximum accuracy of 62.50%. Visualization of the low dimensional manifold of the LI complex via Graph Embedding shows the presence of three distinct stages of LI.

## I. INTRODUCTION

Breast cancer (BC), while a common cancer diagnosis in women, often results in a wide variety of phenotypes. In addition, even within a specific phenotype, prognosis is variable, and the underlying mechanism for this variation is not well understood. Recently, it has been noted that there may be a correlation between the presence of lymphocytic infiltration (LI) and tumor recurrence in breast cancers that have amplification of the HER2 gene (HER2+ BC) [1]. HER2+ breast cancers are mostly high grade cancers that are currently treated with agents that specifically target the HER2 protein. The presence of LI has been observed qualitatively by pathologists on hematoxylin and eosin (H & E) histology specimens of HER2+ BC. Not all HER2+ BC shows presence of LI; however, the identification, quantification, and correlation of LI with genotype and prognosis have proved to be non-trivial

tasks. The stratification of LI obtained via histological image features may in turn correlate with corresponding molecular stratification based on gene expression.

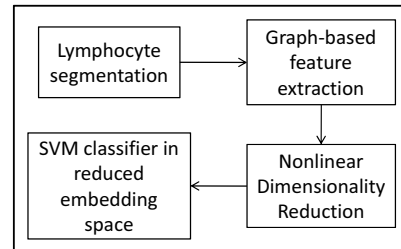


Fig. 1: An overview of the key steps in detection, stratification, and classification of lymphocytes in high grade, HER2+ breast cancer.

The primary objective of this paper is to distinguish between high grade BC studies based on extent of LI. We explore the ability of graph-based architectural features to discriminate between high grade BC studies with varying levels of LI. While texture features (e.g. Varma-Zisserman texon approach [2]) have been applied successfully to the analysis of tissue histology [3], the similarity in textural appearance of LI and BC nuclei suggest that architectural features may be more appropriate. We construct various graphs (Voronoi Diagrams, Delaunay Triangulation, and Minimum Spanning Tree) and extract over 50 architectural features from each histology sample [4], [5]. The graph-based feature are compared against the VZ texon-based classifier, which models grayscale textures as probability distributions of filter responses and distills frequency histograms from the response cluster centers.

A secondary objective lies in the explicit segmentation of lymphocyte required for graph-based feature extraction. Automated discrimination between high grade, HER2+ BC with and without LI on H & E stained biopsy is complicated by the similarity in appearance of cancer nuclei and lymphocytes. We present a method of automated detection that utilizes a Bayesian classifier in conjunction with template matching to detect LI among high grade BC [6]. Our third objective

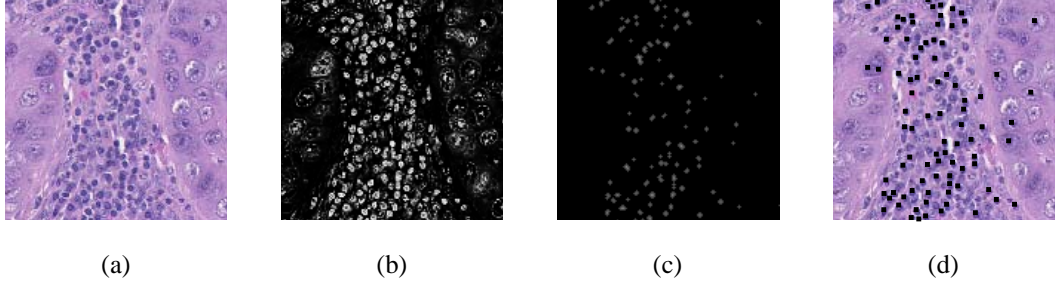


Fig. 2: A (a) breast histology sample with its (b) Bayesian likelihood scene, (c) Euclidean Distance Transform, and (d) with black squares denoting the final segmentation.

stems from the need to visualize and classify high grade BC samples based on their LI content. Since it has been shown that biomedical data is highly nonlinear by nature, we apply a Manifold Learning (ML) technique (Graph Embedding) to reduce the dimensionality of the feature set and avoid the *curse of dimensionality* [7]. The ability to project data into a low-dimensional manifold allows us to discover and visualize the progression of LI. Furthermore, we apply the popular Support Vector Machine (SVM) classifier to the low-dimensional dataset to obtain quantitative classification results.

An overview of our scheme is shown in Figure 1. The rest of the paper is organized as follows. In Section II, we describe the graph-based and VZ feature sets, as well as the automated segmentation, manifold learning, and classification techniques considered. In Section III, we present classification results for each feature set. In Section IV, we present our concluding remarks.

## II. METHODS

### A. Dataset

Anonymized H & E stained breast biopsy samples of HER2+ breast cancers from the Cancer Institute of New Jersey were scanned into a computer using a high resolution whole slide scanner at 20x optical magnification. The images were separated into 3 classes by an expert clinician, comprising 22 with high LI, 10 with medium LI, and 9 with low LI. For the purposes of detection, the samples were reduced to a 2-class scheme by combining the samples with low and medium infiltration. For all methods, we define each image  $\mathcal{C} = (C, g)$  such that  $C$  is a 2D set of pixels  $c \in C$  and the function  $g(c)$  is the image intensity at every  $c \in C$ . Each  $C$  has  $m$  lymphocytes and the centroids of these lymphocytes are defined by  $c_n^1, c_n^2, \dots, c_n^m$ .

### B. Automated Lymphocyte Segmentation

The development of a robust automated lymphocyte segmentation scheme is a vital step for using graph-based features in an unsupervised manner. Separating lymphocytes from surrounding BC nuclei is complicated by the similarities in shape and color between the two structures. Our algorithm takes advantage of the LI's strong hematoxylin staining by using a Bayesian classifier. The relatively small size and

uniformly circular shape of the lymphocyte allows us to use a template matching scheme to discard false positive results.

1) *Training*: Working in the Saturation color channel (in HSV color space), a set of training pixels  $c \in \mathcal{O}_v$  are used to generate a probability distribution function (PDF) for each class  $w_v$ , where  $v \in \{L, B, S\}$  defines the three classes: lymphocytes, BC nuclei, and stroma, respectively. We fit a gamma distribution

$$\mathbf{p}(c \in \mathcal{O}_v; \alpha, \beta) = c^{\alpha-1} \frac{\beta^\alpha e^{-\beta c}}{\Gamma(\alpha)}, \quad (1)$$

where the parameters  $\alpha$  and  $\beta$  are optimized to generate the gamma distribution  $\Gamma$  that best fits  $\mathcal{O}_v$ . This distribution is used as the *a priori* conditional probability  $\mathbf{p}$  to provide uniform probability values as input to the Bayesian classifier.

2) *Low-level intensity information*: The likelihood scene  $\mathcal{L}_v(C)$  is defined as the posterior conditional probability for each pixel  $c \in C$

$$\mathcal{P}(w_v | c, g(c)) = \frac{\mathcal{P}(w_v) \mathbf{p}(c, g(c) | w_v)}{\sum_{i \in \{L, B, S\}} \mathcal{P}(w_i) \mathbf{p}(c, g(c) | w_i)}, \quad (2)$$

where  $\mathcal{P}(w_v)$  represents the non-informative prior probability of each class, and  $\mathbf{p}(c, g(c) | w_v)$  is the *a priori* conditional probability derived from training. The LI likelihood scene  $\mathcal{L}_L$  (Figure 2(b)) is thresholded to create a binary image  $\mathcal{I}_L$ .

3) *Euclidean distance transform (EDT)*: Given  $\mathcal{I}_L$ , the EDT for each foreground pixel is calculated as the distance between that pixel and the nearest background pixel. The EDT is performed on  $\mathcal{I}_L$  to create a gray-level image  $\mathcal{G}_L(C, d)$ . Figure 2(c) shows the corresponding EDT representation for  $\mathcal{I}_L$ .

4) *Template matching*: If the 2D correlation between a circular template and each  $c \in \mathcal{G}_L$  exceeds some empirically-derived threshold, it is included in the foreground of the new binary image  $\mathcal{J}_L$ . The centroids of the resulting foreground regions (Figure 2(d)) in  $\mathcal{J}_L$  are defined as lymphocytes  $\{c_n^1, c_n^2, \dots, c_n^m\}$  for the extraction of graph-based features.

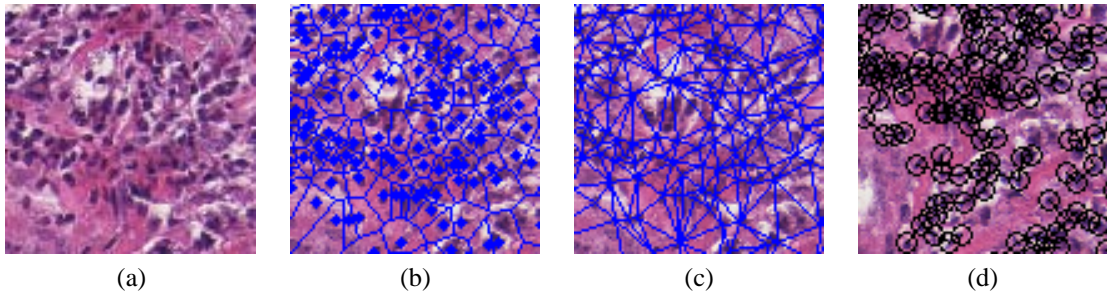


Fig. 3: An (a) histological sample is shown with its (b) Voronoi Diagram, (c) Delaunay Triangulation, and (d) Minimum Spanning Tree displayed over segmented lymphocytes.

### C. Feature Extraction

#### 1) Graph-Based Architectural Features:

**1.1 Voronoi Diagram** - The set of pixels from  $\mathcal{J}_L$  corresponding to segmented lymphocytes  $\{c_n^1, c_n^2, \dots, c_n^m\}$  are used to generate graph-based features [4], [5]. A Voronoi diagram  $\mathbf{V}$  is defined by a set of polygons  $\mathbf{P} = \{P_1, P_2, \dots, P_m\}$ , where a pixel  $c \in C$  is added to polygon  $P_a$  if  $d(c, c_n^a) = \min_j \{\|c - c_n^j\|\}$  where  $a, j \in \{1, 2, \dots, m\}$ . The metric  $d(c, d)$  is defined as the Euclidean distance between two pixels  $c, d \in C$ . The area, perimeter length, and chord length are calculated for all  $\mathbf{P} \in \mathbf{V}$ . The average, standard deviation, min/max ratio and disorder are calculated over all  $\mathbf{P}$ . This provides 12 features for each  $C$ .

**1.2 Delaunay Triangulation** - A Delaunay graph  $\mathbf{D}$  is constructed such that if two unique polygons  $P_a$  and  $P_b$  share a side in  $\mathbf{V}$ , their nuclear centroids  $c_n^a$  and  $c_n^b$ , where  $a, b \in \{1, 2, \dots, m\}$ , are connected by an edge  $E^{a,b}$ . The side lengths and areas are calculated for all triangles in  $\mathbf{D}$  and used to generate the mean, standard deviation, min/max ratio, and disorder values for  $C$ . This provides a total of 10 for each  $C$ .

**1.3 Minimum Spanning Tree** - A spanning tree  $\mathbf{S}$  is a subgraph connecting all vertices of a connected, undirected graph. In each  $\mathbf{S}$ , weights  $\omega_S^E$  are assigned to each edge  $E$  based on the length of  $E$  in  $\mathbf{S}$ . The sum of all weights in  $\mathbf{S}$  determines  $\widehat{\omega}_S$  assigned to each  $\mathbf{S}$ . The MST  $\mathbf{S}^M$  has a weight  $\widehat{\omega}_S^M$  less than or equal to  $\widehat{\omega}_S$  for every other spanning tree  $\mathbf{S}$ . The mean, standard deviation, min/max ratio, and disorder of the branch lengths in  $\mathbf{S}^M$  provide four features for each  $C$ .

**1.4 Nuclear Features** - Nuclear density  $\Pi^D$  is computed as  $\Pi^D = \frac{m}{|C|}$ , where  $|C|$  is the cardinality of  $C$ . Given  $\mathcal{J}_L$  with a set of lymphocytes  $\{c_n^1, c_n^2, \dots, c_n^m\}$ , for each nuclear centroid  $c_n^a$ ,  $N(\zeta, c_n^a)$  is the set of pixels  $c \in C$  contained within a circle with its center at  $c_n^a$  and radius  $\zeta$ . The number of  $c_n^j$ ,  $j \neq a$ ,  $j, a \in \{1, 2, \dots, m\}$  that are in set  $N(\zeta, c_n^a)$  are computed for  $\zeta \in \{10, 20, \dots, 50\}$ . The  $\mu$  required to obtain  $N(\zeta, c_n^a) \in \{3, 5, 7\}$  are also computed. The mean, standard deviation, and disorder of these values for all  $c_n^j$  in  $R$  are calculated to provide 25 features for each  $C$ .

**2) Varma-Zisserman texton-based features:** The texton-based scheme developed by Varma and Zisserman has proved to be a robust approach for textural analysis [2], [3], [8], [9]. The images are converted to grayscale and intensity normalized to provide some invariance to illumination. The Maximum Response 8 (MR8) filter bank contains 38 filters, including edge and bar filters at the orientations  $\theta \in \{0, \pi/6, \dots, 5\pi/6\}$  and scales  $(\sigma_x, \sigma_y) \in \{(3, 1), (6, 2), (12, 4)\}$ . The edge ( $\mathcal{E}$ ) and bar ( $\mathcal{B}$ ) filter kernels are defined at each  $c = (x, y)$  as

$$\mathcal{E}(x, y, \theta) = -\frac{x'}{\sigma_x^2} \left[ \frac{1}{2\pi\sigma_x\sigma_y} e^{-\left(\frac{x'^2}{2\sigma_x^2} + \frac{y'^2}{2\sigma_y^2}\right)} \right], \quad (3)$$

$$\mathcal{B}(x, y, \theta) = \frac{\partial}{\partial x} [\mathcal{E}(x, y, \theta)], \quad (4)$$

where  $x'$  and  $y'$  are defined by the rotation equations  $x' = x \cos(\theta) - y \sin(\theta)$  and  $y' = x \sin(\theta) + y \cos(\theta)$ . The maximum response from the edge and bar filter responses at each scale  $(\sigma_x, \sigma_y)$  over all  $\theta$  is obtained to extract a net total of six responses. The additional two responses are from Gaussian and Laplacian of Gaussian filters, both using  $(\sigma_x, \sigma_y) = (10, 10)$ . Pixels from filter responses of exemplar images are clustered using the  $k$ -means algorithm to generate a texton dictionary. The model of an image  $C$  is defined as the frequency distribution of textons closest (via Euclidean distance metric) to each  $c \in C$  and a Chi-square distance matrix is created. This algorithm is repeated over  $k$ -means clusters  $k \in \{2, 3, 5, 10\}$ . In all instances, the VZ classifier is implemented over filter window sizes  $f \in \{3, 5, 7\}$  so that textural features at different levels are extracted.

#### D. Manifold Learning to visualize underlying structure of LI

Manifold learning (ML) utilizes nonlinear dimensionality reduction techniques (NLDR) to preserve inter- and intra- class relationships from high-dimensional to a lower dimensional representation of the data. It also allows us to visualize the underlying manifold of the data in a low-dimensional embedding space. We apply ML to both the VZ distance matrices and the graph-based architectural feature set using Graph Embedding (GE). Briefly, this method works as follows.

Given two images  $C_p$  and  $C_q$  with feature vectors  $\mathbf{f}_p$  and  $\mathbf{f}_q$  and a total of  $\mathcal{M}$  images, where  $p, q \in \{1, 2, \dots, \mathcal{M}\}$ , GE uses a symmetric  $\mathcal{M} \times \mathcal{M}$  confusion matrix  $\mathcal{S}(p, q) = e^{-\|\mathbf{f}_p - \mathbf{f}_q\|}$  to

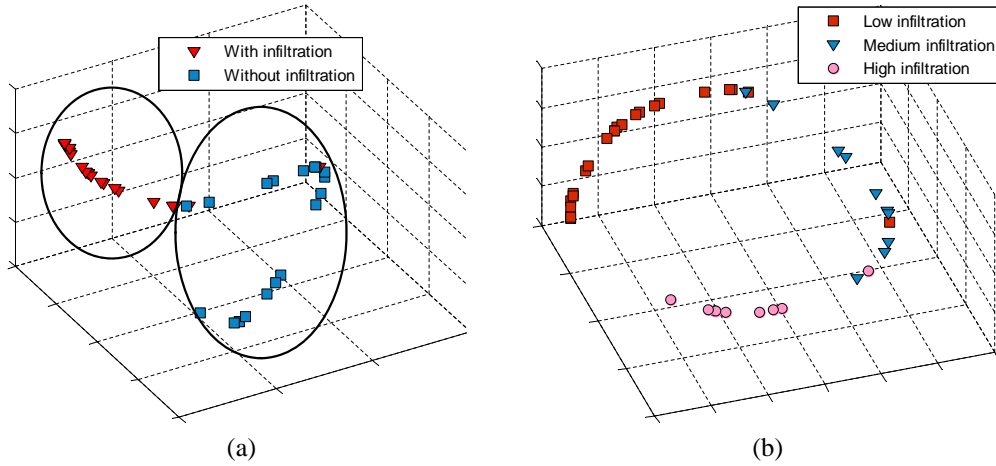


Fig. 4: Scatter plots representing the histological samples (via graph-based features) in reduced three-dimensional GE space. The low-dimensional manifold reveals excellent separation (a) between low and high LI levels, and (b) between 3 classes of LI and also reveals the transition from low to high levels of LI.

find an embedding vector  $\mathcal{X}$ . The optimal embedding vector  $\hat{\mathcal{X}}$  is determined by

$$\hat{\mathcal{X}} = \arg \max_{\mathcal{X}} \left[ 2\eta \frac{\mathcal{X}^T (\mathcal{A} - \mathcal{S}) \mathcal{X}}{\mathcal{X}^T \mathcal{A} \mathcal{X}} \right], \quad (5)$$

where  $\mathcal{A}(p, p) = \sum_q \mathcal{S}(p, q)$  and  $\eta = \mathcal{M} - 1$ . To embed an  $H$ -dimensional dataset into  $h'$ -dimensional space, where  $h' \ll H$ , the optimal embedding for each image  $\hat{\mathcal{X}}(\mathcal{C})$  provides coordinates such that  $\hat{\mathcal{X}}(\mathcal{C}) = [W_z(\mathcal{C}) | z \in \{1, 2, \dots, h'\}]$ , where  $W_z(\mathcal{C})$  are the first  $h'$  eigenvalues associated with  $\mathcal{X}(\mathcal{C})$ .

#### E. Support Vector Machine (SVM) Classifier to test discriminative ability of graph-based features

Training samples from the reduced GE feature space are supplied to a SVM classifier to identify the hyperplane that distinguishes BC studies with low and high levels of infiltration. During testing, a novel sample from the  $h'$  feature space is compared to the hyperplane for classification. Given a dataset of  $M$  samples, the randomized  $\mathcal{N}$ -fold cross-validation technique divides the dataset randomly into  $\mathcal{N} = 3$  groups.  $\mathcal{N} - 1$  groups are used in training the SVM classifier, while the remaining group is used for testing. This process is iterated such that  $\mu_Y$  is the mean accuracy and  $\phi_Y$  is the standard deviation over  $Y = 50$  trials.

### III. RESULTS AND DISCUSSION

#### A. Quantitative Results

Quantitative evaluation via randomized cross-validation suggests that the VZ distance matrices produce a relatively low mean accuracy of  $58.07\% \pm 7.41\%$  over 50 trials (Table I). Conversely, the graph-based features discriminate very well between low and high LI images with an accuracy of  $89.50\% \pm 6.22\%$  over 50 trials.

#### B. Qualitative Results

The low-dimensional graph-based features reveals the underlying structure of the data. The Graph Embedding of the graph-based features reveals the progression from low to high degrees of infiltration on a continuous low-dimensional manifold (Figure 4). This shows the potential of using our scheme for LI grading to determine the degree of LI, rather than simply the presence or absence of LI.

Feature Set	Classification Accuracy
VZ ( $k=2$ )	$50.64\% \pm 6.11\%$
VZ ( $k=3$ )	$60.43\% \pm 7.10\%$
VZ ( $k=5$ )	$61.29\% \pm 8.39\%$
VZ ( $k=10$ )	$59.93\% \pm 8.04\%$
Graph-based	$89.50\% \pm 6.22\%$

TABLE I: Mean classification accuracies from  $Y$ -iteration randomized  $\mathcal{N}$ -fold cross-validation trials with  $\mathcal{N}=3$  and  $Y = 50$ . Results are shown for each VZ distance matrix at  $k = \{2, 3, 5, 10\}$  and the manually segmented graph-based features. All datasets are reduced to 3 dimensions with GE and classified via SVM.

### IV. CONCLUDING REMARKS

The visualization and classification of the presence of LI in HER2+ BC plays an important role in stratifying this subset of high-grade breast cancer. In this paper, we have demonstrated that supervised graph-based (Voronoi, DT, MST, Nuclear) features that exploit the architectural difference in arrangement between cancer nuclei and lymphocytes have greater success compared to the VZ classifier in distinguishing high grade BC based on the presence or absence of LI. Furthermore, the automated segmentation scheme based on a Bayesian classifier, has shown the ability to discriminate LI from BC nuclei in BC histology. The use of an ML scheme (Graph Embedding) permits the visualization of the progression of LI in stages (from low to high). ML may play a significant

role in visualization and understanding of other biological processes related to histopathology. A current limitation of our work is in the tendency of the automated lymphocyte segmentation scheme toward oversegmentation. The approach described here should be broadly applicable to quantifying LI in other tumor types, although some optimization will be required. Future work will focus on improving the robustness of our segmentation scheme and further evaluation on a larger cohort of data for which extensive clinical outcome data exists.

#### V. ACKNOWLEDGEMENTS

This work was made possible due to grants from the Wallace H. Coulter foundation, New Jersey Commission on Cancer Research, the National Cancer Institute (R21CA127186-01, R03CA128081-01), the Society for Imaging and Informatics on Medicine, and the Life Science Commercialization Award. The authors would like to thank Dr. David Foran for his advice, comments, and preparation of the data.

#### REFERENCES

- [1] G. Alexe et al., "High expression of lymphocyte-associated genes in node-negative her2+ breast cancers correlates with lower recurrence rates," *Cancer Res.*, vol. 67, no. 22, pp. 10669–10676, 2007.
- [2] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vision*, vol. 62, no. 1, pp. 61–81, 2005.
- [3] B. Hall et al., "A clinically motivated 2-fold framework for quantifying and classifying immunohistochemically stained specimens," in *MICCAI (2)*, 2007, pp. 287–294.
- [4] S Doyle, M Hwang, K Shah, A Madabhushi, et al., "Automated grading of prostate cancer using architectural and textural image features," *ISBI*, pp. 1284–1287, 2007.
- [5] S. Doyle, A. Madabhushi, et al., "A boosting cascade for automated detection of prostate cancer from digitized histology," in *MICCAI*, 2006, pp. 504–511.
- [6] S. Naik, S. Doyle, A. Madabhushi, et al., "Automated nuclear and gland segmentation and gleason grading of prostate histology by integrating low-, high-level and domain specific information," in *ISBI*, 2008.
- [7] G. Lee, C. Rodriguez, and A. Madabhushi, "Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies," *IEEE ACM T. Comput. Bi.*, vol. 5, no. 3, pp. 368–384, 2008.
- [8] Lin Yang et al., "High throughput analysis of breast cancer specimens on the grid," in *MICCAI (1)*, 2007, pp. 617–625.
- [9] Oncel Tuzel et al., "Classification of hematologic malignancies using texton signatures," *Pattern Anal. Appl.*, vol. 10, no. 4, pp. 277–290, 2007.