

Computerized Image-Based Detection and Grading of Lymphocytic Infiltration in HER2+ Breast Cancer Histopathology

Ajay Nagesh Basavanhally, *Student Member, IEEE*, Shridar Ganesan, Shannon Agner, *Student Member, IEEE*, James Peter Monaco, *Member, IEEE*, Michael D. Feldman, John E. Tomaszewski, Gyan Bhanot, and Anant Madabhushi*, *Senior Member, IEEE*

Abstract—The identification of phenotypic changes in breast cancer (BC) histopathology on account of corresponding molecular changes is of significant clinical importance in predicting disease outcome. One such example is the presence of lymphocytic infiltration (LI) in histopathology, which has been correlated with nodal metastasis and distant recurrence in HER2+ BC patients. In this paper, we present a computer-aided diagnosis (CADx) scheme to automatically detect and grade the extent of LI in digitized HER2+ BC histopathology. Lymphocytes are first automatically detected by a combination of region growing and Markov random field algorithms. Using the centers of individual detected lymphocytes as vertices, three graphs (Voronoi diagram, Delaunay triangulation, and minimum spanning tree) are constructed and a total of 50 image-derived features describing the arrangement of the lymphocytes are extracted from each sample. A nonlinear dimensionality reduction scheme, graph embedding (GE), is then used to project the high-dimensional feature vector into a reduced 3-D embedding space. A support vector machine classifier is used to discriminate samples with high and low LI in the reduced dimensional embedding space. A total of 41 HER2+ hematoxylin-and-eosin-stained images obtained from 12 patients were considered in this study. For more than 100 three-fold cross-validation trials, the architectural feature set successfully distinguished samples of high and low LI levels with a classification accuracy greater than 90%. The popular unsupervised Varma-Zisserman texton-based classification scheme was used for comparison and yielded a classification accuracy of only 60%. Additionally, the projection of the 50 image-derived features for all 41 tissue samples into a reduced dimensional

space via GE allowed for the visualization of a smooth manifold that revealed a continuum between low, intermediate, and high levels of LI. Since it is known that extent of LI in BC biopsy specimens is a prognostic indicator, our CADx scheme will potentially help clinicians determine disease outcome and allow them to make better therapy recommendations for patients with HER2+ BC.

Index Terms—Breast cancer (BC), classification, digital pathology, feature extraction, image analysis, lymphocytic infiltration (LI), nonlinear dimensionality reduction, prognosis, segmentation, texture.

I. INTRODUCTION

BREAST cancer (BC) is the second leading cause of cancer-related deaths in women, with more than 182 000 new cases of invasive BC predicted in the United States for 2008 alone [1]. Although it is a common cancer diagnosis in women, the fact that BC exhibits an exceptionally heterogeneous phenotype in histopathology [2] leads to a variety of prognoses and therapies. One such phenotype is the presence of lymphocytic infiltration (LI) in invasive BC that exhibits amplification of the HER2 gene (HER2+ BC). Most HER2+ BC is currently treated with agents that specifically target the HER2 protein.

Researchers have shown that the presence of LI in histopathology is a viable prognostic indicator for various cancers, including HER2+ BC [3]–[5]. The function of LI as a potential antitumor mechanism in BC was first shown by Aaltomaa *et al.* [4]. More recently, Alexe *et al.* [5] demonstrated a correlation between the presence of high levels of LI and tumor recurrence in early stage HER2+ BC. Consequently, it is surprising that pathologists do not routinely report on the presence of LI, especially in HER2+ BC. A possible reason for this is that pathologists currently lack the automated image analysis tools to accurately, efficiently, and reproducibly quantify the presence and degree of LI in BC histopathology. The ability to automatically detect LI would be invaluable to BC pathologists and oncologists, since manual detection of individual lymphocyte nuclei in BC histopathology is a tedious and time-consuming process, which is not feasible in the clinical setting. The availability of a computerized image analysis scheme for automated quantification of LI extent in HER2+ BC will enable development of an inexpensive image-based system for predicting disease survival and outcome.

While some researchers [6]–[12] have recently begun to develop computer-aided diagnosis (CADx) schemes for the

Manuscript received June 8, 2009; revised August 20, 2009. First published October 30, 2009; current version published February 17, 2010. This work was supported by the Wallace H. Coulter Foundation, the New Jersey Commission on Cancer Research, the National Cancer Institute under Grant R01CA136535-01, Grant ARRA-NCI-3 R21 CA127186-02S1, Grant R21CA127186-01, Grant R03CA128081-01, and Grant R03CA143991-01, the Cancer Institute of New Jersey, and the Life Science Commercialization Award from Rutgers University. Asterisk indicates corresponding author.

A. N. Basavanhally, S. Agner, and J. P. Monaco are with the Department of Biomedical Engineering, Laboratory for Computational Imaging and Bioinformatics, Rutgers University, Piscataway, NJ 08854 USA (e-mail: abasavan@eden.rutgers.edu; agnersc@umdnj.edu; jpm Monaco@rci.rutgers.edu).

S. Ganesan is with the Robert Wood Johnson Medical School, The Cancer Institute of New Jersey, New Brunswick, NJ 08903 USA (e-mail: ganeshas@umdnj.edu).

M. D. Feldman and J. E. Tomaszewski are with the Department of Pathology and Laboratory Medicine, Hospital of the University of Pennsylvania, Philadelphia, PA 19104-6100 USA (e-mail: feldmanm@mail.med.upenn.edu).

G. Bhanot is with the Department of Biology and Biochemistry, and the Department of Physics, Rutgers University, Piscataway, NJ 08854 USA (e-mail: g.bhanot@rci.rutgers.edu).

*A. Madabhushi is with the Department of Biomedical Engineering, Rutgers University, Piscataway, NJ 08854 USA (e-mail: anantm@rci.rutgers.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TBME.2009.2035305

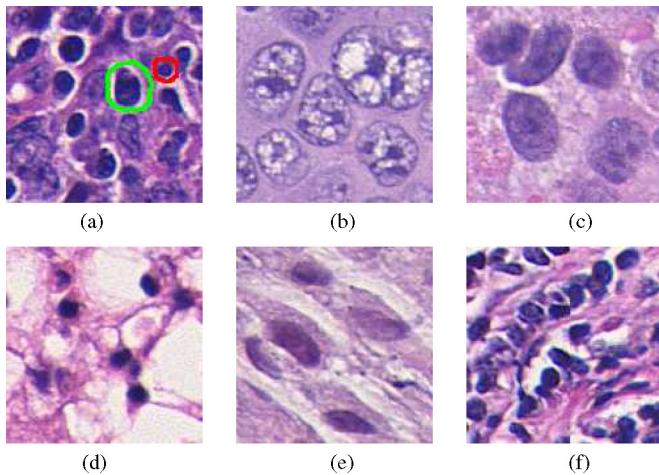


Fig. 1. There are several challenges in automated lymphocyte detection, including (a) the similarity in appearance between a cancer cell nucleus (circled in green) and a lymphocyte nucleus (circled in red). In general, lymphocyte nuclei are distinguished from cancer cell nuclei by their smaller size, more circular shape, and a darker, homogeneous staining. Additional challenges include variations in the appearance of [(b) and (c)] BC nuclei within a single histopathology slide, (d) presence of fat among cancerous tissue, (e) histological fixing, and (f) slide digitization artifacts.

analysis of digitized BC histopathology, they have mostly focused on either finding suspicious regions of interest (ROI) [6], [9] or have attempted to determine cancer grade from manually isolated ROIs [7], [8], [12]. The methods for both applications use image-based features to discriminate between two classes: either normal and benign regions or low- and high-grade ROIs. Specifically, the size and shape of cancer nuclei have been shown to distinguish low- and high-grade histology images [6], [12]. Textural features and filter banks have also been employed [6]–[9], [11] to model the phenotypic appearance of BC histopathology.

An important prerequisite for extracting histopathological image attributes to model BC appearance is the ability to automatically detect and segment histological structures, including nuclei and glands. Consequently, the ability of an image analysis scheme to grade the extent of LI in a BC histopathology image is dependent on the algorithm's ability to automatically detect lymphocytes. Automated lymphocyte detection, however, is a nontrivial task complicated by the intrinsic similarity in appearance of BC nuclei and lymphocyte nuclei on hematoxylin and eosin (H&E) stained breast biopsy samples [see Fig. 1(a)]. In addition, even within a particular slide, the morphology of BC nuclei is highly heterogeneous due to variations in cancer grade and mitotic phase [see Fig. 1(b) and (c)] [13]. Biological differences such as the presence of fat deposits [see Fig. 1(d)] can confound algorithms that rely on boundary detection alone. Preparation issues such as “cutting artifact” [see Fig. 1(e)] and digitization misalignment [see Fig. 1(f)] lead to similar problems, but are more difficult to predict and correct, since they are unrelated to the underlying biology.

While several researchers have been developing algorithms for detection of nuclei [8], [14]–[20] in digitized histopathology, to the best of our knowledge, there have been no attempts to automatically detect or quantify extent of LI on BC histopathology.

Some popular approaches to automated nuclear detection are based on adaptive thresholding [8], [14] and fuzzy c-means clustering [16], [18]. These techniques rely on differences in staining to distinguish nuclei from surrounding tissue. However, they are not appropriate for the task of LI detection due to the similarity in appearance between BC and lymphocyte nuclei [see Fig. 1(a)]. Techniques such as active contours [15], [19], [20] have utilized gradient (edge) information to automatically isolate nuclei in histological images. These methods, however, might be limited in their ability to handle variations in the appearance of BC nuclei [see Fig. 1(b) and (c)] and image acquisition artifacts [see Fig. 1(e) and (f)]. Some researchers have developed hybrid techniques in order to improve nuclear detection and segmentation results. For example, Glotsos *et al.* [19] used support vector machine (SVM) clustering to improve initialization active contour models. More recently, semiautomated probabilistic models have used pixelwise intensity information to detect cancer [17] and lymphocyte nuclei [21] in digitized BC histopathology. Probabilistic models, however, are usually limited by the availability of expert-annotated training data.

To address some of the challenges in automated detection of lymphocytes (as illustrated in Fig. 1), we present a computerized image analysis scheme that combines a region-growing algorithm with maximum *a posteriori* (MAP) estimation and Markov random field (MRF) theory [22]. First, all candidate BC and lymphocyte nuclei are detected via a region-growing algorithm that uses contrast measures to find optimal boundaries [22], [23]. By growing outward from the center of each nucleus, this technique is robust to artifacts outside of the nuclei [see Fig. 1(d)–(f)]. The region-growing algorithm has a high detection sensitivity, resulting in a large number of lymphocyte and nonlymphocyte nuclei being detected. MAP estimation improves detection specificity by incorporating size and luminance information from each detected object to temporarily label it as either a BC or lymphocyte nucleus (these being the two main classes of objects detected). MRF theory [22], [24] then allows us to improve lymphocyte detection specificity by modeling the infiltration phenomenon in terms of spatial proximity, whereby an object is more likely to be labeled as a lymphocyte nucleus if it is surrounded by other lymphocyte nuclei. The application of MRF is a unique step that exploits the spatial properties of LI to: 1) distinguish nuclei that would be, otherwise, misclassified [see Fig. 1(a)] and 2) isolate infiltrating lymphocytes from the surrounding baseline level of lymphocytes. We achieve MAP estimation by using the iterated conditional modes (ICMs) algorithm, a fast and simple method for maximizing the posterior probability that a detected object is indeed a lymphocyte [22], [25].

Detection of lymphocytes alone, however, cannot completely characterize the abnormal LI phenotype because a baseline level of lymphocytes is present in all tissues. Gunduz *et al.* [10] explored automated cancer diagnosis by using hierarchical graphs to model tissue architecture, whereby a graph is defined as a set of vertices (nuclei) with corresponding edges connecting all nuclei. We have previously shown the importance of using graph algorithms to quantitatively describe the spatial arrangement of nuclei in distinguishing cancer grade in both prostate cancer

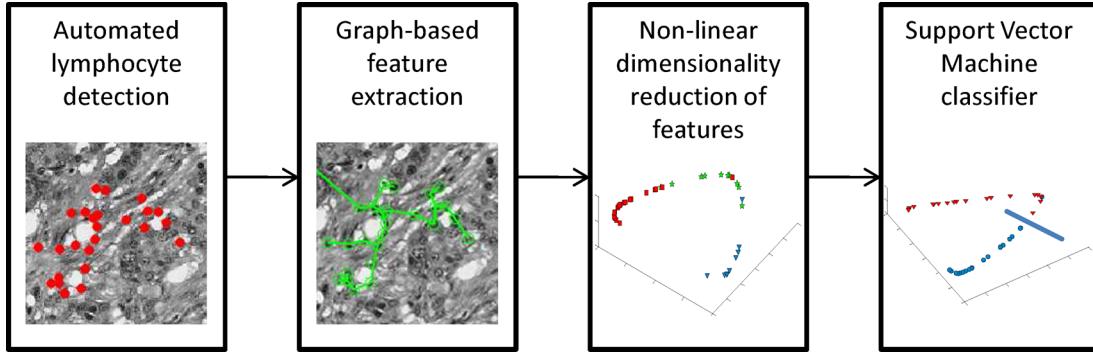


Fig. 2. Flowchart illustrating the four main steps in the CADx scheme for LI-based stratification of HER2+ BC histopathology. Automated lymphocyte detection is followed by feature extraction of architectural and morphological features. The high-dimensional feature space is then nonlinearly embedded into a reduced dimensional space via GE, which allows for data visualization and subsequent evaluation via a SVM classifier.

and BC histopathology [12], [26]. In [12] quantitative features derived from graphs (Voronoi diagram, Delaunay triangulation, and minimum spanning tree) constructed using BC nuclei as vertices were used to successfully stratify low, intermediate, and high BC grade on digitized histopathology. In this paper, we attempt to quantify the extent of LI in HER2+ BC histopathology by extracting graph-based and nuclear features to characterize the architectural arrangement and morphological appearance of lymphocytes [12]. Traditional textural signatures such as first-order gray level features, second-order Haralick statistics, and Gabor filter features were not considered in this paper because they have been shown to be unsuitable for CADx applications in breast and prostate cancer that rely on spatial information [12], [26].

While a large set of descriptive features is certainly desirable for modeling biological processes such as LI, a high-dimensional feature space also presents problems for data classification analysis. First, the curse of dimensionality [27] affects computational efficiency due to the exponential increase in data volume required for each additional feature. Further, it is impossible to directly visualize the relationships between images in a high-dimensional space. We address both issues by utilizing graph embedding (GE) [12], [28], a nonparametric, nonlinear dimensionality reduction scheme [29], to first project the image-derived features onto a reduced 3-D space while simultaneously preserving object-class relationships. Thus, two samples with low levels of LI would be projected relatively close to each other and relatively farther from a sample with a high level of LI.

An SVM classifier [29], [30] is then employed in the reduced dimensional space to discriminate between images with high and low LI. We have previously shown the utility of SVM classifiers in conjunction with architectural and morphological image features to distinguish malignant and benign BC samples, and also for distinguishing between different grades of BC [12] and prostate cancer [26]. SVM results for our architectural features are also compared against results from the popular, unsupervised Varma-Zisserman (VZ) texton-based scheme.

The main components of our methodology for automated detection and stratification of LI on BC histopathology are illustrated in the flowchart in Fig. 2. The organization of the rest of this paper is as follows. First, the dataset of HER2+ BC

histopathology images employed in this study is described in Section II. We then detail the components of the automated lymphocyte detection scheme in Section III. This is followed by a detailed description of the steps involved in image feature extraction (see Section IV). We then describe the dimensionality reduction process (see Section V) used to find a low-dimensional feature space. Methods for evaluating the automated lymphocyte detection scheme, the cross-validation and classification techniques, and the VZ-texon approach are presented in Section VI. Section VII details the results of all quantitative and qualitative experiments. The paper ends with a few concluding remarks and directions for future research (Section VIII).

II. DATASET DESCRIPTION AND NOTATION

A total of 41 H&E stained breast biopsy samples from 12 patients at The Cancer Institute of New Jersey (CINJ) were obtained and scanned into a computer using a high-resolution whole-slide scanner at $20\times$ optical magnification ($0.33\text{ }\mu\text{m}$ spatial resolution). The size of each image falls within $600 \leq U_X \leq 700$ and $500 \leq U_Y \leq 600$, where U_X and U_Y are the width and height, respectively, in pixels. These images were separated into three classes by a BC oncologist based on the extent of LI. The dataset comprises 22 low, ten medium, and nine high LI samples. For the purpose of quantitative classification (as described in Section VI-B), the oncologist separated the images into two classes comprising 22 low LI and 19 high LI samples, respectively.

We define a dataset $\mathbf{Z} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_{\mathcal{M}}\}$ of \mathcal{M} images. An image scene $\mathcal{C} \in \mathbf{Z}$ is defined as $\mathcal{C} = (C, g)$, where C is a 2-D set of pixels $c \in C$ and g is the associated luminance function from the International Commission on Illumination (CIE) laboratory color space [31]. A list of symbols and notation commonly used in this paper is shown in Table I. The CIE-laboratory color space has the advantage of being more perceptually uniform, and more robust to variations in staining and digitization than red-green-blue (RGB) space [see Fig. 1 (a)] [9], [31].

III. AUTOMATED NUCLEAR DETECTION OF LYMPHOCYTES

Beginning with a set of N candidate lymphocyte nuclear centers $\mathbf{N} = \{n_1, n_2, \dots, n_N\}$, we attempt to identify a set of L finalized lymphocyte nuclei with centers given by $\mathcal{O} =$

TABLE I
LIST OF KEY NOTATION

Symbol	Description
$Z = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_M\}$	HER2+ BC histopathology dataset comprising M digitized images
$\mathcal{C} = (C, g)$	Image scene defined by a set of pixels (C) and luminance function (g)
$N = \{n_1, n_2, \dots, n_N\}$	N candidate lymphocyte nuclei centers in image scene \mathcal{C}
$R = \{r_1, r_2, \dots, r_N\}$	N candidate regions grown from N
$O = \{o_1, o_2, \dots, o_L\}$	L finalized lymphocyte nuclei centers in image scene \mathcal{C} , where $O \subseteq N$
R	Set of pixels representing lymphocyte nucleus region S_{CR}^*
$X_r \in \{\omega_b, \omega_\ell\}$	Random variable denoting class BC (ω_b) or lymphocyte (ω_ℓ) nucleus for each region $r \in R$
$Y_r = [A_r, \sigma_r]^\top \in \mathbb{R}^{+2}$	Random variable denoting features square root of area (A) and std. dev. intensity (σ) for each region $r \in R$
x_r, y_r	Specific instances of X_r and Y_r
\mathbf{x}, \mathbf{y}	Sets of $x_r, \forall r \in R$ and $y_r, \forall r \in R$
$G = \{\mathcal{O}, \mathbf{E}, \mathbf{W}\}$	Graph with vertex-set \mathcal{O} , edge-set \mathbf{E} , and weights \mathbf{W}
$\mathbf{F}(\mathcal{C})$	Architectural feature set for image scene \mathcal{C}
$\mathbf{F}'(\mathcal{C})$	Low-dimensional embedding of architectural feature set for image scene \mathcal{C}
$\mathcal{Y}(\mathcal{C}) \in \{+1, -1\}$	True label for image scene \mathcal{C} as determined by expert pathologist, such that $+1$ represents high LI and -1 represents low LI.

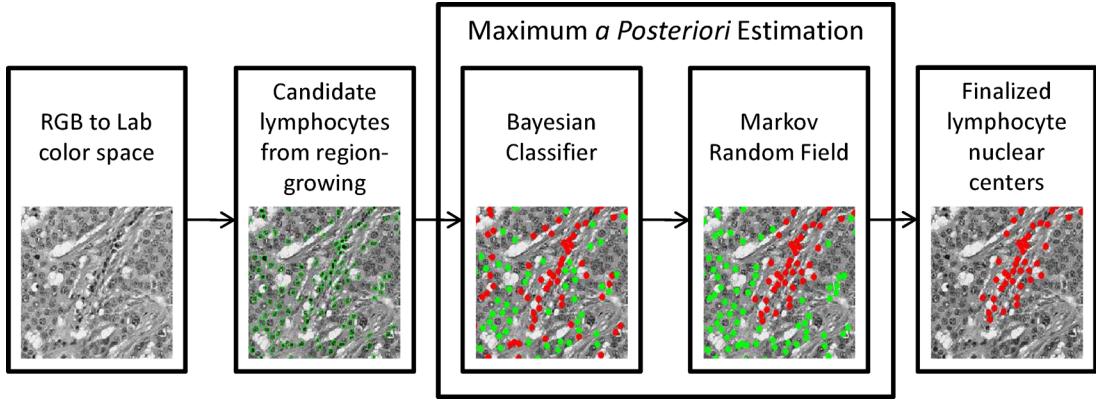


Fig. 3. Flowchart illustrating the main steps in the automated lymphocyte detection scheme.

$\{o_1, o_2, \dots, o_L\}$, such that $O \subseteq N$. The following sections detail the region-growing, MAP estimation, and MRF algorithms that comprise the lymphocyte detection module of our CADx system (see Fig. 3).

A. Candidate Lymphocyte Detection via Region Growing

We first attempt to identify candidate image locations that could represent centers of lymphocytic nuclei. The region-growing algorithm exploits the fact that lymphocyte nuclei in the luminance channel are identified as continuous, circular regions of low intensity, circumscribed by sharp, well-defined boundaries (see Fig. 1). The image scene \mathcal{C} is convolved with a Gaussian (smoothing) kernel at multiple scales $\sigma_G \in \{6, 7, 8\} \mu\text{m}$ to account for variations in lymphocyte size. After convolution at each scale, valleys (i.e., the darkest pixels) are found on the smoothed image based on local differences in luminance. These valleys define a set of seed points $N = \{n_1, n_2, \dots, n_N\}$ that represent candidate lymphocyte centers on the original scene \mathcal{C} . Each $n \in N$ is grown into a corresponding region $r \in R$ using the four-step procedure described next. See [23] for additional details on the region-growing scheme.

- Step 1: A set of current pixels $S_{CR} = \{n\}$ is initialized, as shown in Fig. 4(a). The current boundary S_{CB} is defined as the set of eight-connected pixels surrounding S_{CR} . A

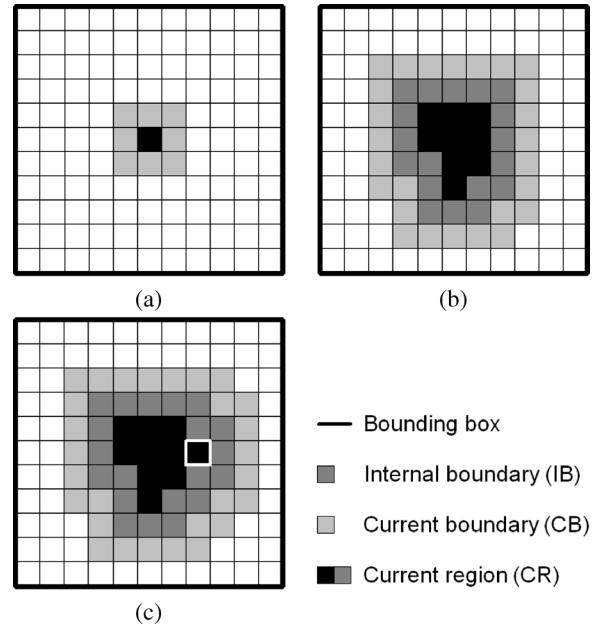


Fig. 4. Schematic illustrating the iterative growth of a region r . After initialization of the current region S_{CR} (a) current boundary S_{CB} , and bounding box S_{BB} , (b) new pixels are added iteratively. (c) When a new pixel (outlined in white) is added to S_{CR} , the boundaries S_{CB} and S_{IB} are adjusted accordingly.

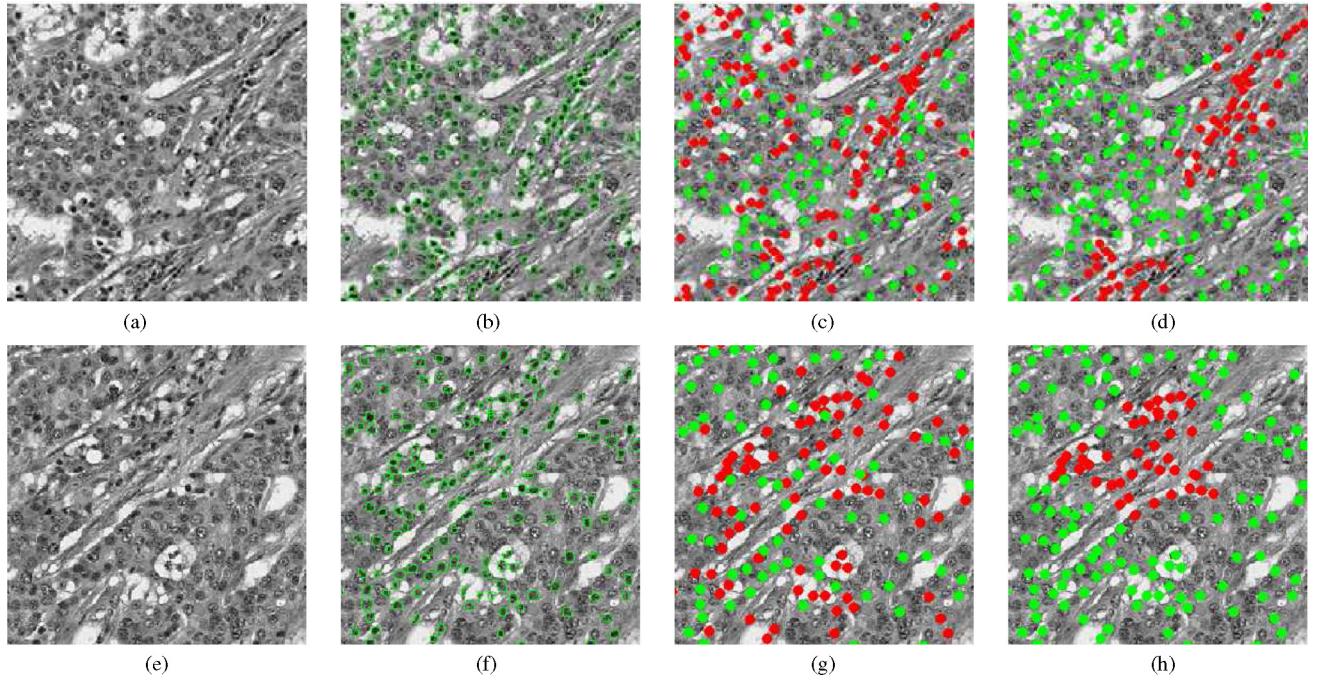


Fig. 5. (a) and (e) Luminance channels of two different HER2+ BC histopathology studies and corresponding results for [(b) and (f)] initial region-growing-based lymphocyte detection. (c) and (g) Preliminary Bayesian refinement showing detected BC nuclei in green and detected lymphocyte nuclei in red. (d) and (h) Final lymphocyte detection result after the MRF pruning step.

square bounding box S_{BB} containing all pixels within a $12\sigma_G \times 12\sigma_G$ neighborhood around n is then constructed.

- 2) *Step 2:* The pixel $c \in S_{\text{CB}}$ with the lowest intensity in the current boundary is identified. Pixel c is removed from S_{CB} and added to S_{CR} . The current boundary S_{CB} is updated to include every pixel $d \in S_{\text{BB}}$ that is an eight-connected neighbor of c and $d \notin S_{\text{CR}}$. A set of internal boundary pixels $S_{\text{IB}} \subset S_{\text{CR}}$ [see Fig. 4(b) and (c)] is defined as all pixels in S_{CR} that are eight-connected to any pixel in S_{CB} .
- 3) *Step 3:* The variables \bar{g}_{IB} and \bar{g}_{CB} are computed as the mean intensity of pixels in S_{IB} and S_{CB} , respectively. The boundary strength is computed at each iteration as $\bar{g}_{\text{IB}} - \bar{g}_{\text{CB}}$.
- 4) *Step 4:* Steps 2 and 3 are iterated until the current region S_{CR} tries to add a pixel outside the bounding box S_{BB} . The optimal lymphocyte region S_{CR}^* is identified at the iteration for which the boundary strength $\bar{g}_{\text{IB}} - \bar{g}_{\text{CB}}$ is maximum [see Fig. 5(b) and (f)].

Since the region-growing procedure is repeated with seed points from a variety of smoothing scales $\sigma_G \in \{6, 7, 8\} \mu\text{m}$, overlapping regions are resolved by discarding the region with the lower boundary strength. For the sake of convenience, we will refer to S_{CR}^* as R in the rest of this paper.

B. Bayesian Modeling of LI via MAP Estimation

The initial lymphocyte detection is refined by incorporating domain knowledge regarding lymphocyte size, luminance, and spatial proximity. Each $r \in \mathbf{R}$ has two associated random variables: $X_r \in \Lambda \equiv \{\omega_b, \omega_\ell\}$ indicating its classification as either a

BC (ω_b) or lymphocyte (ω_ℓ) nucleus and $Y_r \equiv [A_r, \sigma_r]^\top \in \mathbb{R}^{+2}$ denoting the two observed features

$$A_r = \sqrt{|R|} \quad (1)$$

$$\sigma_r = \sqrt{\frac{1}{|R|} \sum_{c \in R} (g(c) - \bar{g})^2} \quad (2)$$

where A_r is the square root of nuclear area (1), σ_r is the standard deviation of luminance in the nuclear region (2), $|R|$ is the cardinality of R , and $\bar{g} = (1/|R|) \sum_{c \in R} g(c)$ is the average pixel intensity of R . The choice of the two features (A_r and σ_r) is motivated by the fact that: 1) BC nuclei are typically larger than lymphocyte nuclei and 2) BC and lymphocyte nuclei are significantly different in terms of the homogeneity of their luminance values. Specific instances of the random variables X_r and Y_r are denoted by $x_r \in \Lambda$ and $y_r = [A_r, \sigma_r]^\top \in \mathbb{R}^{+2}$, respectively. We define the random variables collectively for all $r \in \mathbf{R}$ as $\mathbf{X} = \{X_1, X_2, \dots, X_N\}$ and $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_N\}$ with state spaces $\Omega = \Lambda^N$ and $\mathbb{R}^{+2 \times N}$, respectively. Instances of \mathbf{X} and \mathbf{Y} are denoted by variables $\mathbf{x} = (x_1, x_2, \dots, x_N) \in \Omega$ and $\mathbf{y} = (y_1, y_2, \dots, y_N) \in \mathbb{R}^{+2 \times N}$.

The labels $\mathbf{X} = \mathbf{x}$, given the feature vectors $\mathbf{Y} = \mathbf{y}$, are estimated using MAP estimation [32], which advocates finding the \mathbf{x} that maximizes the posterior probability

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \quad (3)$$

where $p(\mathbf{y}|\mathbf{x})$ is the likelihood term and $p(\mathbf{x})$ and $p(\mathbf{y})$ are prior distributions for \mathbf{x} and \mathbf{y} , respectively. Since maximization of (3) is only with respect to \mathbf{x} , the prior distribution $p(\mathbf{y})$ is ignored.

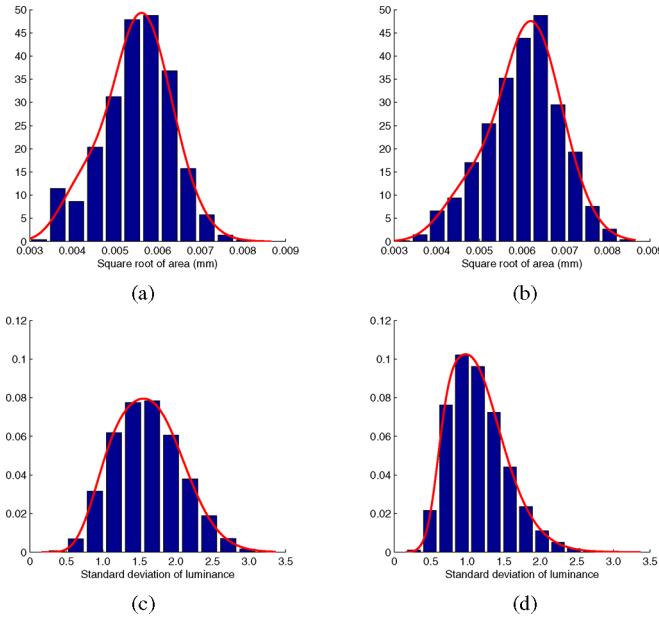


Fig. 6. PDFs estimated from empirical training data and modeled via weighted sum of gamma distributions for [(a) and (c)] ω_ℓ and [(b) and (d)] ω_b classes for [(a) and (b)] square root of area and [(c) and (d)] variance in luminance of each $r \in \mathbf{R}$. In each distribution [(a)–(d)], the estimated parametric model is overlaid.

1) Modeling Lymphocyte Features via Trained Probability Distributions: The likelihood term $p(\mathbf{y}|\mathbf{x})$ in (3) is calculated from probability density functions (PDFs), where \mathbf{x} is provided by manual delineation of lymphocytes in a training set. Under the assumption that \mathbf{y} is independent and identically distributed, the likelihood term in (3) can be simplified such that

$$p(\mathbf{y}|\mathbf{x}) = \prod_{r \in \mathbf{R}} p(y_r|x_r). \quad (4)$$

Each 2-D PDF is modeled as the product of two independent distributions: $p(y_r|x_r) = \mathcal{F}(A_r|x_r)\mathcal{F}(\sigma_r|x_r)$. Thus, we require four 1-D PDFs $\mathcal{F}(A_r|\omega_b)$, $\mathcal{F}(A_r|\omega_\ell)$, $\mathcal{F}(\sigma_r|\omega_b)$, and $\mathcal{F}(\sigma_r|\omega_\ell)$, as shown in Fig. 6. To reduce local irregularities and create a smooth, continuous distribution, the 1-D PDFs are modeled by mixtures of Gamma distributions [33]

$$\bar{\Gamma}(z; \delta, \phi, t) = \delta z^{t_1-1} \frac{e^{-z/\phi_1}}{\phi_1^{t_1} \Gamma(t_1)} + (1-\delta) z^{t_2-1} \frac{e^{-z/\phi_2}}{\phi_2^{t_2} \Gamma(t_2)} \quad (5)$$

where $z \in \mathbb{R}^+$, $\delta \in [0, 1]$ is the mixing parameter, $t_1, t_2 > 0$ are the shape parameters, $\phi_1\phi_2 > 0$ are the scale parameters, and Γ is the Gamma function [33]. Calculating $p(\mathbf{y}|\mathbf{x})$ allows us to estimate (3) and assign a tentative class $x_r \in \{\omega_b, \omega_\ell\}$ to each $r \in \mathbf{R}$ [see Fig. 5(c) and (g)].

2) Modeling Lymphocyte Proximity via MRFs: The prior distribution $p(\mathbf{x})$ [see (3)] is defined by an MRF. The Markov property [24] states that

$$p(x_r|\mathbf{x}_{-r}) = p(x_r|\mathbf{x}_{\eta_r}) \quad (6)$$

where the neighborhood η_r is empirically assumed to contain all regions within a $30 \mu\text{m}$ radius of r , $\mathbf{x}_{-r} = \{x_s : s \in \mathbf{R}, s \neq r\}$, and $\mathbf{x}_{\eta_r} = \{x_s : s \in \eta_r\}$. We use the ICMs algorithm [25], a

deterministic relaxation procedure, to perform MAP estimation [see (3)] and assign a hard label $x_r \in \{\omega_b, \omega_\ell\}$ to each $r \in \mathbf{R}$. Thus, each region is classified as either a BC or lymphocyte nucleus [see Fig. 5(d) and (h)]. The regions labeled as BC nuclei are discarded, while centers of the L lymphocyte nuclei regions are computed and stored as $\mathcal{O} = \{o_1, o_2, \dots, o_L\}$.

IV. FEATURE EXTRACTION

We define the complete, undirected graph $\mathcal{G} = (\mathcal{O}, \mathbf{E}, \mathbf{W})$, where $\mathcal{O} = \{o_1, o_2, \dots, o_L\}$ is the set of vertices corresponding to the set of lymphocyte nuclear centroids, $\mathbf{E} = \{E_1, E_2, \dots, E_m\}$ is the set of edges connecting the nuclear centroids such that $\{(o_i, o_j) \in \mathbf{E} : \forall o_i, o_j \in \mathcal{O}, i, j \in \{1, 2, \dots, L\}, i \neq j\}$, and $\mathbf{W} = \{W_1, W_2, \dots, W_m\}$ is a set of weights proportional to the length of each $E \in \mathbf{E}$. To extract information about the arrangement of lymphocyte nuclei, we construct subgraphs representing the Voronoi diagram \mathcal{G}_V , Delaunay triangulation \mathcal{G}_D , and minimum spanning tree \mathcal{G}_{MST} (see Fig. 7). In addition, statistics describing the number and density of nuclei are calculated directly from \mathcal{O} .

A. Voronoi Diagram

The Voronoi graph $\mathcal{G}_V = (\mathcal{O}, \mathbf{E}_V, \mathbf{W}_V)$ [see Fig. 7(b) and (f)] is a spanning subgraph of \mathcal{G} defined as a set of polygons $\mathbf{P} = \{P_1, P_2, \dots, P_L\}$ surrounding all nuclear centroids \mathcal{O} [34]. Each pixel $c \in C$ is linked with the nearest centroid $o \in \mathcal{O}$ (via Euclidean distance) and added to the associated polygon $P \in \mathbf{P}$. The mean, standard deviation, minimum/maximum (min/max) ratio, and disorder (i.e., standard deviation divided by the mean) are calculated for the area, perimeter length, and chord length over all \mathbf{P} , yielding a set of 13 features (\mathbf{f}_V) for each scene C (see Table II).

B. Delaunay Triangulation

The Delaunay graph $\mathcal{G}_D = (\mathcal{O}, \mathbf{E}_D, \mathbf{W}_D)$ [see Fig. 7(c) and (g)] is a spanning subgraph of \mathcal{G} and the dual graph of \mathcal{G}_V [34]. It is constructed such that if $P_i, P_j \in \mathbf{P}$ share a side, where $i, j \in \{1, 2, \dots, L\}$, and their nuclear centroids $o_i, o_j \in \mathcal{O}$ are connected by an edge $(o_i, o_j) \in \mathbf{E}_D$. The mean, standard deviation, min/max ratio, and disorder are calculated for the side length and area of all triangles in \mathcal{G}_D , yielding a set of eight features (\mathbf{f}_D) for each scene C (see Table II).

C. Minimum Spanning Tree

A spanning tree $\mathcal{G}_S = (\mathcal{O}, \mathbf{E}_S, \mathbf{W}_S)$ refers to any spanning subgraph of \mathcal{G} [34]. The total weight $\widehat{\mathbf{W}}_S$ for each subgraph is determined by summing all individual weights $W \in \mathbf{W}_S$. The minimum spanning tree \mathcal{G}_{MST} [see Fig. 7(d) and (h)] is the spanning tree with the lowest total weight such that $\mathcal{G}_{MST} = \arg \min_{\mathcal{G}_S \in \mathcal{G}} [\widehat{\mathbf{W}}_S]$. The mean, standard deviation, min/max ratio, and disorder of the branch lengths in \mathcal{G}_{MST} yield a set of four features (\mathbf{f}_{MST}) for each scene C (see Table II).

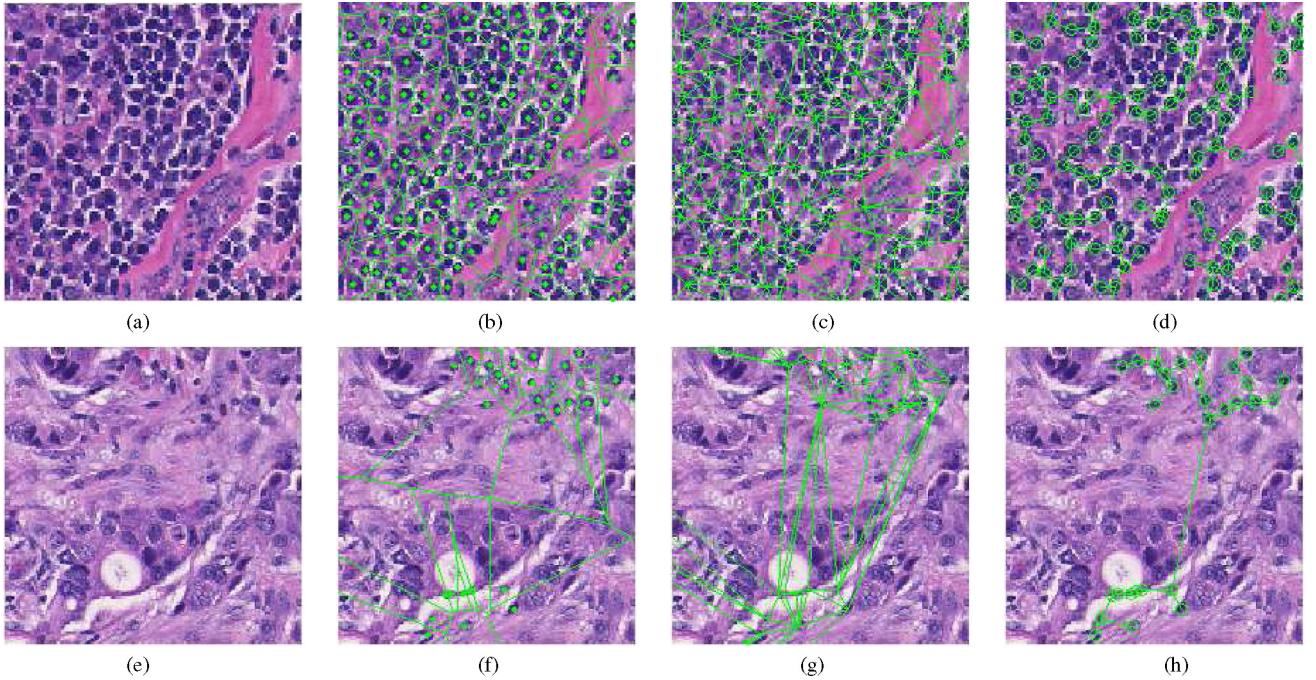


Fig. 7. Two different HER2+ BC histopathology images with (a) high and (b) low levels of LI. (b) and (f) Corresponding Voronoi diagrams constructed using the automatically detected lymphocyte centers as vertices of the graph. (c) and (g) Corresponding Delaunay triangulation and [(d) and (h)] minimum spanning tree graphs.

TABLE II
BREAKDOWN OF 50 ARCHITECTURAL FEATURES, COMPRISING 25
GRAPH-BASED AND 25 NUCLEAR ATTRIBUTES

Feature Set	Description	No. of features
f_V	Total area of all polygons	13
	Polygon area: mean, std dev., min/max ratio, disorder	
	Polygon perimeter: mean, std dev., min/max ratio, disorder	
	Polygon chord length: mean, std dev., min/max ratio, disorder	
f_D	Triangle side length: mean, std dev., min/max ratio, disorder	8
	Triangle area: mean, std dev., min/max ratio, disorder	
f_{MST}	Edge length: mean, std dev., min/max ratio, disorder	4
f_{NF}	Density of nuclei	25
	Distance to $\{3, 5, 7\}$ nearest nuclei: mean, std dev., disorder	
	Nuclei in $\zeta \in \{10, 20, \dots, 50\}$ pixel radius: mean, std dev., disorder	

D. Nuclear Features

The global density $L/|C|$ of lymphocyte nuclei is calculated for each scene \mathcal{C} , where L is the total number of detected lymphocytes and $|C|$ represents the number of pixels (cardinality) in \mathcal{C} . For any nuclear centroid $o_i \in \mathcal{O}$, we define a corresponding nuclear neighborhood $\eta^\zeta(o_i) = \{o_j : \|o_i - o_j\|_2 < \zeta, o_j \in \mathcal{O}, o_j \neq o_i\}$, where $\zeta \in \{10, 20, \dots, 50\}$ and $\|\cdot\|_2$ is the L2 norm. The mean, standard deviation, and disorder of $\eta^\zeta(o_i)$, $\forall o_i \in \mathcal{O}$ are calculated. Additionally, we estimate the minimum radius ζ^* such that $|\eta^{\zeta^*}(o_i)| \in \{3, 5, 7\}$, and calculate the mean, standard deviation, and disorder over all $o_i \in \mathcal{O}$.

A total of 25 nuclear features (f_{NF}) are extracted for each scene \mathcal{C} (see Table II).

V. NONLINEAR DIMENSIONALITY REDUCTION VIA GE

GE is employed to nonlinearly transform the high-dimensional set of image features into a low-dimensional embedding while preserving relative distances between images from the original feature space [12]. For each scene \mathcal{C} , a 50-D image feature set is defined as the superset $\mathbf{F} = \{f_V, f_D, f_{MST}, f_{NF}\}$ containing all features derived from the Voronoi diagram, Delaunay triangulation, minimum spanning tree, and nuclear statistics. Given histopathology images \mathcal{C}_a and \mathcal{C}_b with corresponding image feature sets $\mathbf{F}(\mathcal{C}_a)$ and $\mathbf{F}(\mathcal{C}_b)$, where $a, b \in \{1, 2, \dots, \mathcal{M}\}$, a $\mathcal{M} \times \mathcal{M}$ confusion matrix $\mathcal{W}_{\mathbf{F}}(a, b) = \exp(-\|\mathbf{F}(\mathcal{C}_a) - \mathbf{F}(\mathcal{C}_b)\|_2) \in \mathbb{R}^{\mathcal{M} \times \mathcal{M}}$ is constructed. The optimal embedding vector \mathbf{F}' is obtained from the maximization of the following function:

$$\mathcal{E}(\mathbf{F}') = 2(\mathcal{M} - 1)\text{trace} \left[\frac{\mathbf{F}'^\top (\mathcal{A} - \mathcal{W}_{\mathbf{F}})\mathbf{F}'}{\mathbf{F}'^\top \mathcal{A} \mathbf{F}'} \right] \quad (7)$$

where \mathcal{A} is a diagonal matrix defined $\forall a \in \{1, 2, \dots, \mathcal{M}\}$ as $\mathcal{A}(a, a) = \sum_b [\mathcal{W}_{\mathbf{F}}(a, b)]$. The lower dimensional embedding space is defined by the eigenvectors corresponding to the β smallest Eigen values of $(\mathcal{A} - \mathcal{W}_{\mathbf{F}})\mathbf{F}' = \lambda \mathcal{A} \mathbf{F}'$. In this paper, an empirically determined value of $\beta = 3$ was used. The matrix $\mathbf{F}'(\mathbf{Z}) \in \mathbb{R}^{\mathcal{M} \times \beta}$ of the first β eigenvectors is constructed such that $\mathbf{F}'(\mathbf{Z}) = \{\mathbf{F}'(\mathcal{C}_1), \mathbf{F}'(\mathcal{C}_2), \dots, \mathbf{F}'(\mathcal{C}_{\mathcal{M}})\}$.

VI. EVALUATION METHODS

A. Hausdorff Distance for Evaluation of Automated Lymphocyte Detection Scheme

The automated lymphocyte detection scheme is evaluated by the Hausdorff distance [35], a similarity measure used to compare the fidelity of automated detection against the “gold standard” obtained by manual inspection. For each image scene \mathcal{C} , lymphocyte centroids from the automated ($v \in \mathcal{O}^{\text{auto}}$) and manual ($u \in \mathcal{O}^{\text{man}}$) detection schemes are identified. The centroid locations in \mathcal{O}^{man} were estimated exhaustively by an expert pathologist who manually labeled the individual lymphocytes in each scene. The partial, directed Hausdorff distance is calculated for $\mathcal{O}^{\text{auto}}$ with respect to \mathcal{O}^{man} as

$$\Phi_H(\mathcal{O}^{\text{auto}}, \mathcal{O}^{\text{man}}) = \min_{u \in \mathcal{O}^{\text{man}}} \|v - u\|_2, \forall v \in \mathcal{O}^{\text{auto}}. \quad (8)$$

B. Cross-Validation Using SVM Classifier for Quantitative Evaluation of Architectural Features

The SVM classifier [30] is employed to evaluate the ability of the image descriptors to discriminate between high and low levels of LI in HER2+ BC histopathology images. We construct the SVM classifier by using a Gaussian kernel function Π to project training data $\mathbf{Z}_{\text{tra}} \subset \mathbf{Z}$ onto a higher dimensional space. This high-dimensional representation allows the SVM to construct a hyperplane to separate the two classes (i.e., high and low LI). The classifier is then evaluated by projecting testing data $\mathbf{Z}_{\text{tes}} \subset \mathbf{Z}$ into the same space and recording the locations of the newly embedded samples with respect to the hyperplane.

Given BC histopathology images $\mathcal{C}_a, \mathcal{C}_b \in \mathbf{Z}_{\text{tra}}$ with corresponding low-dimensional embedding vectors $\mathbf{F}'(\mathcal{C}_a)$ and $\mathbf{F}'(\mathcal{C}_b)$, $a, b \in \{1, 2, \dots, M\}$, respectively, the Gaussian kernel $\Pi(\mathbf{F}'(\mathcal{C}_a), \mathbf{F}'(\mathcal{C}_b)) = \exp(-\epsilon (\|\mathbf{F}'(\mathcal{C}_a) - \mathbf{F}'(\mathcal{C}_b)\|_2)^2)$, where ϵ is a scaling factor that normalizes $\mathbf{F}'(\mathcal{C}_a)$ and $\mathbf{F}'(\mathcal{C}_b)$, is used to project the data into the high-dimensional SVM space [17]. The general form of the SVM is given as

$$\Theta(\mathcal{C}_a) = \sum_{\gamma=1}^{\tau} \xi_{\gamma} \mathcal{Y}(\mathcal{C}_{\gamma}) \Pi(\mathbf{F}'(\mathcal{C}_a), \mathbf{F}'(\mathcal{C}_{\gamma})) + \mathbf{b} \quad (9)$$

where $\gamma \in \{1, 2, \dots, \tau\}$ represents the τ marginal training samples (i.e., support vectors), \mathbf{b} is the hyperplane bias estimated for \mathbf{Z}_{tra} , and ξ_{γ} is the model parameter determined by maximizing the objective function [29], [30]. The true image label $\mathcal{Y}(\mathcal{C}_b) \in \{+1, -1\}$ represents a high or low level of LI, as determined by an expert pathologist. The output of the SVM classifier, $\Theta(\mathcal{C}_a)$, represents the distance from image scene \mathcal{C}_a to the hyperplane. A testing image scene $\mathcal{C}_a \in \mathbf{Z}_{\text{tes}}$ is determined to be classified correctly if $\mathcal{Y}(\mathcal{C}_a) = \text{sign}[\Theta(\mathcal{C}_a)]$.

The Gaussian kernel Π has recently become popular for classification in a number of biomedical image processing applications [17], [36]. In this paper, we use the Gaussian kernel instead of the traditional linear kernel [29] because its nonlinear projection creates additional separation between the data points in the high-dimensional SVM space, and hence, simplifies the classification task.

One problem with the SVM classifier is that it is susceptible to bias from the arbitrary selection of training and testing samples [32]. A k -fold cross-validation scheme [32] is used to mitigate this bias by selecting training samples in a randomized manner and running the SVM classifier multiple times. First, \mathbf{Z} is divided randomly into k subsets, while ensuring that images from each class $\mathcal{Y} \in \{+1, -1\}$, are proportionally represented in each of the k subsets. Since our objective was binary classification, the samples labeled as medium LI were aggregated with the low LI images. Hence, the goal of the SVM classifier was to distinguish 22 high and 19 low LI samples. All samples from $k - 1$ subsets are pooled together to obtain \mathbf{Z}_{tra} and the remaining subset is used as \mathbf{Z}_{tes} . For each of the k iterations, an SVM classifier is trained with \mathbf{Z}_{tra} and evaluated on \mathbf{Z}_{tes} ; a new \mathbf{Z}_{tes} and \mathbf{Z}_{tra} being chosen at each iteration so that all samples are evaluated. Using a value of $k = 3$, the entire cross-validation algorithm was repeated over 100 trials, and the resulting mean (μ_{ACC}) and standard deviation (σ_{ACC}) of the classification accuracy were obtained. Classification accuracy is defined as the ratio between the number of correctly classified images and the total number of images in the dataset.

C. Formulation of VZ Texton-Based Classifier

The performance of the architectural feature set was compared against the VZ texton-based features [37] for distinguishing between the 41 low and high LI images in \mathbf{Z} . Textons have previously been shown to be useful in applications related to content-based image retrieval [38] and computer-aided classification [39] for digitized cancer histopathology.

- 1) *Step 1:* All $\mathcal{C}_{\text{tra}} \in \mathbf{Z}_{\text{tra}}$ are first convolved with the maximum response 8 (MR8) filter bank [37], which contains edge and bar filters at several orientations and scales. An 8-D MR8 feature vector $\mathbf{f}_{\text{text}}(c)$ is defined for each $c \in C, \forall \mathcal{C}_{\text{tra}} \in \mathbf{Z}_{\text{tra}}$.
- 2) *Step 2:* Feature vectors \mathbf{f}_{text} of all $c \in C, \forall \mathcal{C}_{\text{tra}} \in \mathbf{Z}_{\text{tra}}$ are clustered using the K -means algorithm [32], and the K cluster centers $\{c_1^*, c_2^*, \dots, c_K^*\}$ are defined as textons.
- 3) *Step 3:* For each $c \in \mathcal{C}_{\text{tra}}$, the closest corresponding texton $c_j^*, j \in \{1, 2, \dots, K\}$ is identified based on $\arg \min_j \|\mathbf{f}_{\text{text}}(c) - \mathbf{f}_{\text{text}}(c_j^*)\|_2$. A texton histogram is constructed for each $\mathcal{C}_{\text{tra}} \in \mathbf{Z}_{\text{tra}}$ as $\mathcal{H}(\mathcal{C}_{\text{tra}}) = (\mathbf{H}, h)$, where \mathbf{H} is a 1-D grid of K bins and $h(j)$ represents the number of $c \in \mathcal{C}_{\text{tra}}$ identified as being closer to c_j^* than any other texton.
- 4) *Step 4:* For each novel image scene $\mathcal{C}_{\text{tes}} \in \mathbf{Z}_{\text{tes}}$, a corresponding texton histogram $\mathcal{H}(\mathcal{C}_{\text{tes}})$ is computed. The training image scene $\mathcal{C}_{\text{tra}}^* \in \mathbf{Z}_{\text{tra}}$ that is most similar to \mathcal{C}_{tes} is identified based on

$$\mathcal{C}_{\text{tra}}^* = \arg \min_{\mathcal{C}_{\text{tra}} \in \mathbf{Z}_{\text{tra}}} [\chi^2(\mathcal{H}(\mathcal{C}_{\text{tra}}), \mathcal{H}(\mathcal{C}_{\text{tes}}))] \quad (10)$$

where $\chi^2(\mathcal{H}(\mathcal{C}_{\text{tra}}), \mathcal{H}(\mathcal{C}_{\text{tes}}))$ is the Chi-squared distance [40] between the histograms of \mathcal{C}_{tra} and \mathcal{C}_{tes} . If $\mathcal{Y}(\mathcal{C}_{\text{tes}}) = \mathcal{Y}(\mathcal{C}_{\text{tra}}^*)$, \mathcal{C}_{tes} is said to have been correctly classified; otherwise, incorrectly classified. For additional details on the VZ texton approach, see [37].

TABLE III

RESULTS OF SVM CLASSIFICATION ACCURACY (μ_{ACC} , σ_{ACC}) FOR 41 BC HISTOPATHOLOGY IMAGES USING 100 THREE-FOLD CROSS-VALIDATION TRIALS FOR AUTOMATED AND MANUAL LYMPHOCYTE DETECTION WITH ARCHITECTURAL (BOTH REDUCED \mathbf{F}' AND UNREDUCED \mathbf{F}) AND VZ IMAGE FEATURES

Feature Set	Classification Accuracy (%)
$\mathbf{F}'(\mathbf{Z})$ (automated detection)	90.41 ± 2.97
$\mathbf{F}'(\mathbf{Z})$ (manual detection)	94.59 ± 1.72
$\mathbf{F}(\mathbf{Z})$ (automated detection)	89.71 ± 2.83
$\mathbf{F}(\mathbf{Z})$ (manual detection)	99.59 ± 0.92
VZ ($K = 2$)	48.17 ± 6.08
VZ ($K = 3$)	60.20 ± 5.66
VZ ($K = 5$)	58.63 ± 7.17
VZ ($K = 10$)	56.17 ± 7.63

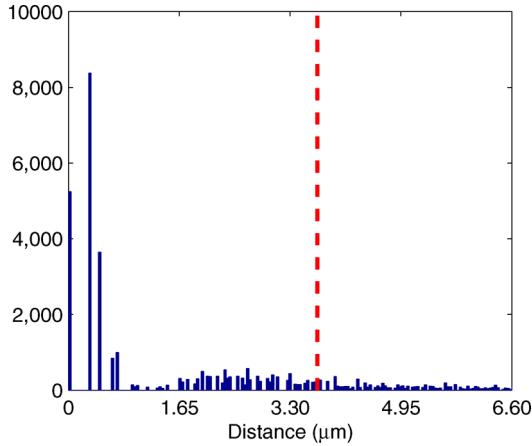


Fig. 8. Histogram of the partial, directed Hausdorff distances $\Phi_H(\mathcal{O}^{\text{auto}}, \mathcal{O}^{\text{man}})$ between automatically and manually detected lymphocyte nuclei in all 41 BC histopathology images. The red dashed line denotes the median of the errors of the automated lymphocyte detection scheme.

The mean μ_{ACC} and standard deviation σ_{ACC} of the classification accuracy of the VZ-texton approach are calculated over 100 randomized three-fold cross-validation trials (see Table III). These experiments were repeated for each $K \in \{2, 3, 5, 10\}$, producing a maximum classification accuracy of $60.20\% \pm 5.66\%$ (see Table III).

VII. RESULTS AND DISCUSSION

A. Performance of Automated Lymphocyte Detection Scheme

Over a total of $|\mathcal{O}^{\text{auto}}| = 42\,000$ automatically detected lymphocyte nuclei for all $C \in \mathbf{Z}$, the median partial Hausdorff distance was determined to be $3.70 \mu\text{m}$ (see Fig. 8). Considering an average lymphocyte diameter of approximately $7 \mu\text{m}$, these results verify the ability of our detection scheme to accurately detect LI in HER2+ BC histopathology imagery. Furthermore, the validity of the detection scheme is implicitly borne out in the quantitative classification results discussed in Section VII-B and Table III.

B. Performance of Graph-Based and Nuclear Features

Table III shows the classification accuracies of the reduced feature set $\mathbf{F}'(\mathbf{Z})$ resulting from both automated and manual lymphocyte detection via the SVM classifier. Note that the

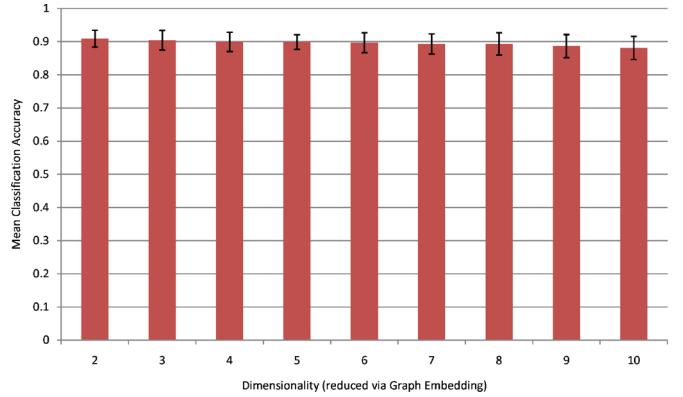


Fig. 9. Mean (μ_{ACC}) classification accuracy over 100 trials of three-fold cross-validation is shown for different dimensionalities $\{2, \dots, 10\}$ obtained via GE. The error bars represent standard deviation (σ_{ACC}) of the classification accuracy.

classification accuracies and variances obtained from the automated detection ($90.41\% \pm 2.97\%$) and manual detection ($94.59\% \pm 1.72\%$) schemes are comparable, reflecting the efficacy of our automated detection scheme. Table III also reveals that the unreduced architectural features $\mathbf{F}(\mathbf{Z})$ (via automated lymphocyte detection) achieve a classification accuracy of $89.71\% \pm 2.83\%$, suggesting, in turn, that GE does not lead to any significant loss in class discriminatory information.

In order to determine the optimal dimensionality for performing classification, the architectural feature set $\mathbf{F}(\mathbf{Z})$ was reduced to various dimensionalities $\{2, 3, \dots, 10\}$ via GE. For each dimensionality, the corresponding μ_{ACC} and error bars (σ_{ACC}) over 100 trials of randomized three-fold cross-validation were calculated (see Fig. 9). Fig. 9 suggests that classification accuracy is stable at lower dimensionalities and drops off slightly at higher dimensionalities.

C. Performance of VZ Features

The classification results (see Table III) show that the VZ textural features did not perform, as well as the architectural features, in distinguishing between BC histopathology samples with high and low levels of LI, with a maximum classification accuracy of $60.20\% \pm 5.66\%$. This result suggests that texture descriptors are unable to quantitatively describe phenotypic changes due to variation in LI extent. Furthermore, both natural variations in histology and imperfections arising from slide preparation (see Fig. 1) may have adversely affected the performance of textural features, since the dataset was not screened to exclude such samples. Conversely, architectural features remain unaffected by these issues because they exploit intrinsic properties such as lymphocyte size, shape, intensity, and arrangement to classify the BC histopathology images.

D. Low-Dimensional Manifold Visualization

Apart from its help to deal with the curse of dimensionality problem for classification, another important application of GE is in its ability to help visualize the underlying structure of the data. Fig. 10 shows the reduced dimensional representation

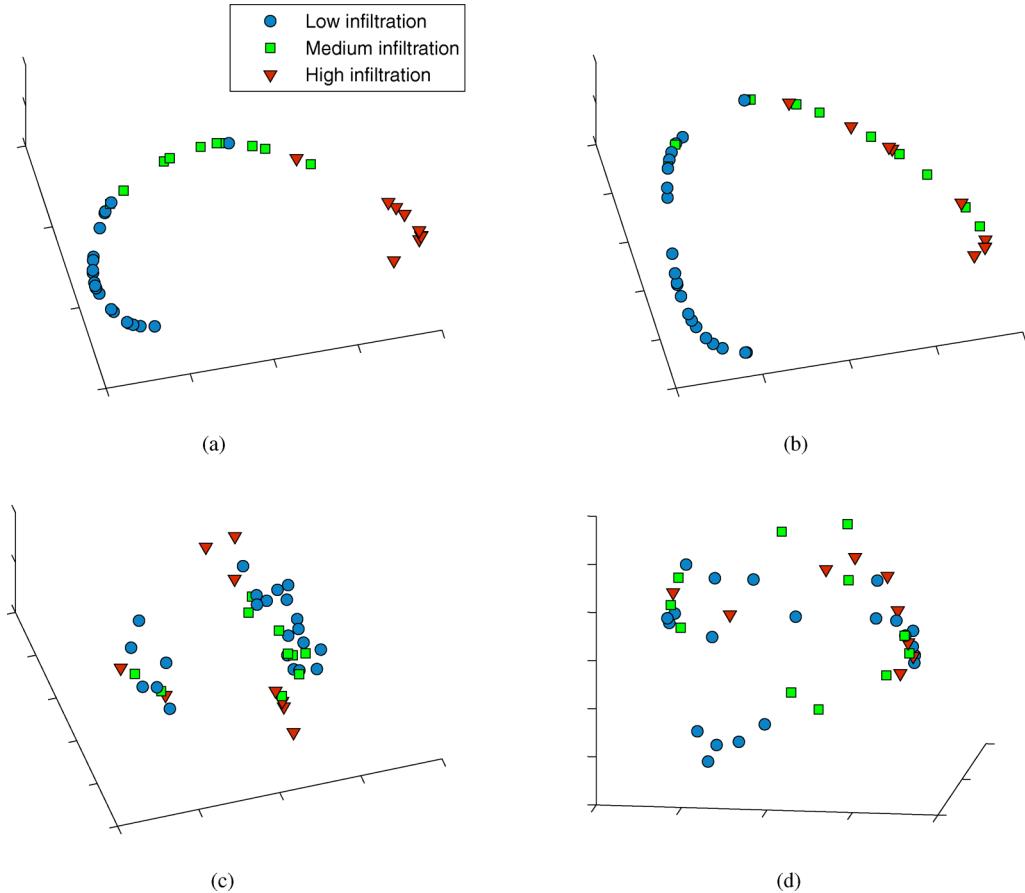


Fig. 10. All 41 images plotted in the GE reduced 3-D eigenspace for the architectural feature set derived from (a) manual and (b) automated lymphocyte detection. Embeddings of the VZ features are shown for (c) $K = 3$ and (d) $K = 5$. The labels denote samples with low LI (blue circles), medium LI (green squares), and high LI (red triangles), as determined by an expert oncologist. Note that GE with the architectural features reveals the presence of an underlying manifold structure showing a smooth continuum of BC samples with low, medium, and high levels of LI.

(three dimensions) of the high dimensional architectural and VZ-texture feature spaces. Note that the three axes in each of Fig. 10(a)–(d) reflect the principal eigenvectors obtained embedding the data via GE. Reducing the architectural feature set to three dimensions via GE reveals the progression from low to medium to high degrees of LI on a smooth, continuous manifold [see Fig. 10(a) and (b)]. Conversely, the VZ features [see Fig. 10(c) and (d)] neither produce a continuous manifold, nor appear to stratify samples based on LI extent. The plots in Fig. 10 further validate the quantitative classification results shown in Table III and reflect the efficacy of architectural image features in stratifying extent of LI.

VIII. CONCLUSION

The primary clinical objective of this paper is to develop a quantitative CADx system for detecting and stratifying the extent of LI in digitized HER2+ BC histopathology images. In this paper, we have demonstrated the ability to automatically detect and grade the extent of LI using architectural features. Additionally, nonlinearly reducing the high-dimensional architectural image feature space reveals the presence of a smooth, continuous manifold on which BC samples with progressively increasing LI are arranged in a continuum. The region-growing

algorithm and subsequent MRF-based refinement allow us to isolate LI from the surrounding BC nuclei, stroma, and baseline level of lymphocytes. The architectural (graph-based and nuclear) features, which exploit differences in arrangement of LI, were found to be more successful than textural (VZ) features in distinguishing LI extent. While applying GE to the high-dimensional feature space did not adversely affect the classification accuracy of the SVM classifier, in conjunction with the architectural and morphological features, it did allow for the visualization of a smooth data manifold. A similar manifold was not reproducible with the VZ features, reflecting that the architectural and morphological features accurately captured class-discriminatory information regarding the spatial extent of LI. The LI classification results were comparable for automated and manual detection, reflecting the robustness of our detection algorithm. The ability of our image analysis algorithm to stratify the extent of LI into low, medium, and high grades will have significant translational and prognostic significance, and could be developed into a prognostic test for predicting disease survival and patient outcome. Furthermore, since the methods comprising our CADx scheme are generalizable, our image classifier could be employed to characterize LI extent in other tissues and diseases.

REFERENCES

- [1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun, "Cancer statistics, 2008," *CA Cancer J. Clin.*, vol. 58, no. 2, pp. 71–96, 2008.
- [2] F. Bertucci and D. Birnbaum, "Reasons for breast cancer heterogeneity," *J. Biol.*, vol. 7, no. 6, 2008.
- [3] J. R. van Nagell, E. S. Donaldson, E. G. Wood, and J. C. Parker, "The significance of vascular invasion and lymphocytic infiltration in invasive cervical cancer," *Cancer*, vol. 41, no. 1, pp. 228–234, Jan. 1978.
- [4] S. Aaltomaa, P. Lippinen, M. Eskelinen, V. M. Kosma, S. Marin, E. Alhava, and K. Syrjanen, "Lymphocyte infiltrates as a prognostic variable in female breast cancer," *Eur. J. Cancer*, vol. 28A, no. 4/5, pp. 859–864, 1992.
- [5] G. Alexe, G. S. Dalgin, D. Scanfeld, P. Tamayo, J. P. Mesirov, C. DeLisi, L. Harris, N. Barnard, M. Martel, A. J. Levine, S. Ganesan, and G. Bhanot, "High expression of lymphocyte-associated genes in node negative her2+ breast cancers correlates with lower recurrence rates," *Cancer Res.*, vol. 67, no. 22, pp. 10 669–10 676, Nov. 2007.
- [6] W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian, "Computer-derived nuclear features distinguish malignant from benign breast cytology," *Hum. Pathol.*, vol. 26, no. 7, pp. 792–796, Jul. 1995.
- [7] B. Weyn, G. van de Wouwer, A. van Daele, P. Scheunders, D. van Dyck, E. van Marck, and W. Jacob, "Automated breast tumor diagnosis and grading based on wavelet chromatin texture description," *Cytometry*, vol. 33, no. 1, pp. 32–40, Sep. 1998.
- [8] S. Petushi, F. U. Garcia, M. M. Haber, C. Katsinis, and A. Tozeren, "Large-scale computations on histology images reveal grade differentiating parameters for breast cancer," *BMC Med. Imag.*, vol. 6, no. 14, 2006.
- [9] B. Karaali and A. Tzeren, "Automated detection of regions of interest for tissue microarray experiments: An image texture analysis," *BMC Med. Imag.*, vol. 7, no. 2, 2007.
- [10] C. Gunduz, B. Yener, and S. H. Gultekin, "The cell graphs of cancer," *Bioinformatics*, vol. 20, pp. i145–i151, Aug. 2004.
- [11] B. H. Hall, M. Ianosi-Irimie, P. Javidian, W. Chen, S. Ganesan, and D. J. Foran, "Computer-assisted assessment of the human epidermal growth factor receptor 2 immunohistochemical assay in imaged histologic sections using membrane isolation algorithm and quantitative analysis of positive controls," *BMC Med. Imag.*, vol. 8, no. 11, 2008.
- [12] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *Proc. 5th IEEE Int. Symp. Biomed. Imag.: From Nano to Macro*, 2008, pp. 496–499.
- [13] S. J. Shin, E. Hyjek, E. Early, and D. M. Knowles, "Intratumoral heterogeneity of her-2/neu in invasive mammary carcinomas using fluorescence in-situ hybridization and tissue microarray," *Int. J. Surg. Pathol.*, vol. 14, no. 4, pp. 279–284, Oct. 2006.
- [14] F. Schnorrenberg, C. Pattichis, K. Kyriacou, and C. Schizas, "Computer-aided detection of breast cancer nuclei," *IEEE Trans. Inf. Technol. Biomed.*, vol. 1, no. 2, pp. 128–140, Jun. 1997.
- [15] P. Bamford and B. Lovell, "Unsupervised cell nucleus segmentation with active contours," *Signal Process.*, vol. 71, no. 2, pp. 203–213, 1998.
- [16] L. Latson, B. Sebek, and K. A. Powell, "Automated cell nuclear segmentation in color images of hematoxylin and eosin-stained breast biopsy," *Anal. Quant. Cytol. Histol.*, vol. 25, no. 6, pp. 321–331, Dec. 2003.
- [17] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *Proc. 5th IEEE Int. Symp. Biomed. Imag.: From Nano to Macro (ISBI)*, 2008, pp. 284–287.
- [18] W. H. Land, D. W. McKee, T. Zhukov, D. Song, and W. Qian, "A kernelised fuzzy-support vector machine CAD system for the diagnosis of lung cancer from tissue images," *Int. J. Funct. Inf. Pers. Med.*, vol. 1, pp. 26–52, 2008.
- [19] D. Glotsos, P. Spyridonos, D. Cavouras, P. Ravazoula, P.-A. Dadioti, and G. Nikiforidis, "Automated segmentation of routinely hematoxylin-eosin-stained microscopic images by combining support vector machine clustering and active contour models," *Anal. Quant. Cytol. Histol.*, vol. 26, no. 6, pp. 331–340, Dec. 2004.
- [20] H. Fatakdawala, A. Basavanhally, J. Xu, G. Bhanot, S. Ganesan, M. Feldman, J. Tomaszewski, and A. Madabhushi, "Expectation maximization driven geodesic active contour: Application to lymphocyte segmentation on digitized breast cancer histopathology," in *Proc. IEEE Conf. Bioinf. Bioeng. (BIBE)*, 2009, pp. 69–76.
- [21] A. Basavanhally, S. Agner, G. Alexe, G. Bhanot, S. Ganesan, and A. Madabhushi, (2008). Manifold learning with graph-based features for identifying extent of lymphocytic infiltration from high grade, her2+ breast cancer histology. presented at the Workshop Microsc. Image Anal. Appl. Biol. (in Conjunction MICCAI), New York [Online]. Available: <http://www.miaab.org/miaab-2008-papers/27-miaab-2008-paper-21.pdf>
- [22] J. P. Monaco, J. E. Tomaszewski, M. D. Feldman, M. Moradi, P. Mousavi, A. Boag, C. Davidson, P. Abolmaesumi, and A. Madabhushi, (2008). Detection of prostate cancer from whole-mount histology images using Markov random fields. presented at the Workshop Microsc. Image Anal. Appl. Biol. (in conjunction MICCAI), New York [Online]. Available: <http://www.miaab.org/miaab-2008-papers/28-miaab-2008-paper-22.pdf>
- [23] S. Hojjatoleslami and J. Kittler, "Region growing: A new approach," *IEEE Trans. Image Process.*, vol. 7, no. 7, pp. 1079–1084, Jul. 1998.
- [24] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.
- [25] J. Besag, "Statistical analysis of dirty pictures," *J. R. Statist. Soc.*, vol. 48, no. 3, pp. 259–302, 1986.
- [26] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of prostate cancer using architectural and textural image features," in *Proc. 4th IEEE Int. Symp. Biomed. Imag.: From Nano to Macro*, 2007, pp. 1284–1287.
- [27] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton Univ. Press, 1961.
- [28] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
- [29] G. Lee, C. Rodriguez, and A. Madabhushi, "Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 5, no. 3, pp. 368–384, Jul.–Sep. 2008.
- [30] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [31] M. W. Schwarz, W. B. Cowan, and J. C. Beatty, "An experimental comparison of rgb, yiq, lab, hsv, and opponent color models," *ACM Trans. Graph.*, vol. 6, no. 2, pp. 123–158, 1987.
- [32] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York: Wiley, 2001.
- [33] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1965.
- [34] J. Sudbo, R. Marcelpoil, and A. Reith, "New algorithms based on the Voronoi diagram applied in a pilot study on normal mucosa and carcinomas," *Anal. Cell. Pathol.*, vol. 21, no. 2, pp. 71–86, 2000.
- [35] D. Huttenlocher, G. Klanderman, and W. Ruckridge, "Comparing images using the hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.
- [36] I. El Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, (Dec. 2000). A support vector machine approach for detection of microcalcifications. *IEEE Trans. Med. Imag.* [Online]. 21(12), pp. 1552–1563. Available: <http://dx.doi.org/10.1109/TMI.2002.806569>
- [37] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vis.*, vol. 62, no. 1/2, pp. 61–81, 2005.
- [38] F. Schnorrenberg, C. S. Pattichis, C. N. Schizas, and K. Kyriacou, "Content-based retrieval of breast cancer biopsy slides," *Technol. Health Care*, vol. 8, no. 5, pp. 291–297, 2000.
- [39] O. Tuzel, L. Yang, P. Meer, and D. J. Foran, "Classification of hematologic malignancies using texton signatures," *Pattern Anal. Appl.*, vol. 10, no. 4, pp. 277–290, 2007.
- [40] J. Naik, S. Doyle, A. Basavanhally, S. Ganesan, M. Feldman, J. Tomaszewski, and A. Madabhushi, (2009). A boosted distance metric: Application to content based image retrieval and classification of digitized histopathology. *SPIE Med. Imag.* [Online]. 7260, pp. 72603F-1–72603F-12. Available: <http://dx.doi.org/10.1117/12.813931>



Ajay Nagesh Basavanhally (S'09) received the B.S. degree in biomedical engineering from Case Western Reserve University, Cleveland, OH, in 2007.

He is currently a Graduate Assistant in the Laboratory for Computational Imaging and Bioinformatics, Department of Biomedical Engineering, Rutgers University, Piscataway, NJ, where he is engaged in the development of computer-aided diagnosis and prognosis tools for breast cancer histopathology.



Shridar Ganesan received the A.B. degree in chemistry from Princeton University, Princeton, NJ, in 1985, and the Ph.D. degree in cell biology and the M.D. degree from Yale University, New Haven, CT, in 1993.

He was a Trainee in internal medicine at Brigham and Womens Hospital. He was a Medical Oncology Fellow at the Dana-Farber Cancer Institute. He is currently an Assistant Professor of medicine at Robert Wood Johnson Medical School, The Cancer Institute of New Jersey, New Brunswick, NJ. His research interests include molecular characterization of breast cancer subtypes.



John E. Tomaszewski received the M.D. degree from the University of Pennsylvania, Philadelphia.

He is a Professor and the Interim Chair of Pathology and Laboratory Medicine at The University of Pennsylvania, Philadelphia. He is a nationally recognized expert in diagnostic genitourinary pathology. His current research interests include high-resolution MRI of prostate and breast cancer, computer-assisted diagnosis, and high-dimensionality data fusion in the creation of a new diagnostic testing paradigms.



Shannon Agner (S'09) received the A.B. degree in engineering sciences and the M.S. degree in engineering from Dartmouth College, Hanover, NH, in 2002 and 2004, respectively. She is currently working toward the M.D. and Ph.D. degrees at the University of Medicine and Dentistry, New Jersey-Robert Wood Johnson Medical School program at Rutgers University, Piscataway, NJ.

She is currently a Predoctoral Student in the Laboratory for Computational Imaging and Bioinformatics, Rutgers University.

Ms. Agner is the recipient of the New Jersey Commission on Cancer Research Predoctoral Fellowship for her research in computer-aided diagnosis of breast cancer on Dynamic contrast enhanced MRI.



Gyan Bhanot received the Ph.D. degree in theoretical particle physics from Cornell University, Ithaca, NY, in 1979.

He was a Postdoctoral Researcher at Brookhaven National Laboratory, The European Center for Nuclear Physics, Geneva, Switzerland, the Institute for Advanced Study, Princeton, and the Institute for Theoretical Physics, Santa Barbara. He was an Associate Professor in the Physics Department, Florida State University. From 1994 to 2006, he was a Research Staff Member in the Physics Department, IBM Research, where he was involved in various projects involving high-performance computing, culminating in a Gordon Bell Award for application performance on the IBM Blue Gene/L, the fastest computer in the world at the time. He is currently a Professor in both the Department of Biology and Biochemistry, and the Department of Physics at Rutgers University, Piscataway, NJ. He is a member of the Cancer Institute of New Jersey and a long-term Visiting Member at the Simons Center for Systems Biology, Institute for Advanced Study in Princeton.



James Peter Monaco (S'02–M'07) received the B.S. degree in computer engineering and the M.S. degree in electrical engineering from Texas A&M University, College Station, in 1996 and 1998, respectively, and the Ph.D. degree in electrical engineering from the University of Texas, Austin, in 2007.

He was with Texas Instruments, Inc., and Raytheon Corporation, where he was engaged in designing IR tracking systems for military aircraft. For six years, he was with VuCOMP, a medical imaging company dedicated to the automatic detection of disease. He is currently a Postdoctoral Fellow in the Laboratory for Computational Imaging and Bioinformatics, Rutgers University, Piscataway, NJ.



Anant Madabhushi (S'98–M'04–SM'09) received the B.S. degree in biomedical engineering from Mumbai University, Mumbai, India, in 1998, the M.S. degree in biomedical engineering from the University of Texas, Austin, in 2000, and the Ph.D. degree in bioengineering from the University of Pennsylvania, Philadelphia, in 2004.

Since 2005, he has been an Assistant Professor in the Department of Biomedical Engineering, Rutgers University, Piscataway, NJ, where he is also the Director of the Laboratory for Computational Imaging and Bioinformatics. He is also a member of the Cancer Institute of New Jersey and an Adjunct Assistant Professor of Radiology at the Robert Wood Johnson Medical Center, New Brunswick, NJ. He has more than 60 publications and book chapters in leading international journals and peer-reviewed conferences, and several patents pending in the areas of medical image analysis, computer-aided diagnosis, machine learning, and computer vision.



Michael D. Feldman received the M.D. and Ph.D. degrees from the University of Medicine and Dentistry of New Jersey.

He is an Associate Professor in the Department of Pathology and Laboratory Medicine at the Hospital of the University of Pennsylvania, Philadelphia, where he is also the Assistant Dean for Information Technology and the Medical Director of Pathology Informatics. His current research interests include the development, integration, and adoption of information technologies in the discipline of pathology, especially in the field of digital imaging.

Dr. Madabhushi is the recipient of a number of awards for both research as well as teaching, including the Coulter Phase 1 and Phase 2 Early Career Award (2006 and 2008), the Excellence in Teaching Award (2007–2009), the Society for Imaging Informatics in Medicine New Investigator Award (2008), and the Life Sciences Commercialization Award (2008).