AUTOMATED IMAGE-BASED DETECTION AND GRADING OF LYMPHOCYTIC INFILTRATION IN BREAST CANCER HISTOPATHOLOGY

by

AJAY BASAVANHALLY

A thesis submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

and

The Graduate School of Biomedical Sciences

University of Medicine and Dentistry of New Jersey

in partial fulfillment of the requirements for the

Joint Degree of Master of Science

Graduate Program in Biomedical Engineering

Written under the direction of

Dr. Anant Madabhushi

and approved by

_____

_____

_____

New Brunswick, New Jersey

Jan, 2010

ABSTRACT OF THE THESIS

Automated Image-based Detection and Grading of Lymphocytic Infiltration in Breast Cancer Histopathology

By Ajay Basavanhally

Thesis Director:
Dr. Anant Madabhushi

The identification of phenotypic changes in breast cancer (BC) histopathology is of significant clinical importance in predicting disease outcome and prescribing appropriate therapy. One such example is the presence of lymphocytic infiltration (LI) in histopathology, which has been correlated with a variety of prognoses and theragnoses (i.e. response to treatment) in BC patients. In this thesis work a computer-aided diagnosis (CADx) system is detailed for quantitatively measuring the extent of LI from hematoxylin and eosin (H & E) stained histopathology. The CADx system is subsequently applied to BC patients expressing the HER2 gene (HER2+ BC), where LI extent has been found to correlate with nodal metastasis and distant recurrence. Although LI may be graded qualitatively by BC pathologists, there is currently no quantitative and reproducible method for measuring LI extent in HER2+ BC histopathology. Hence, a CADx system that performs this task will potentially help clinicians predict disease outcome and allow them to make better therapy recommendations for HER2+ BC patients. The CADx methodology comprises three

key steps. First, a combination of region-growing and Markov Random Field algorithms is used to detect individual lymphocyte nuclei and isolate areas of LI in digitized H & E stained histopathology images. The centers of individual detected lymphocytes are used as vertices to construct a series of graphs (Voronoi Diagram, Delaunay Triangulation, and Minimum Spanning Tree) and a total of 50 architectural features describing the spatial arrangement of lymphocytes are extracted from each image. By using Graph Embedding, a non-linear dimensionality reduction method, to project the high-dimensional feature vectors into a reduced 3D embedding space, it is possible to visualize the underlying manifold that represents the continuous nature of the LI phenotype. Over a set of 100 randomized cross-validation trials, a Support Vector Machine classifier shows that the architectural feature set distinguishes HER2+ BC histopathology samples containing high and low levels of LI with a classification accuracy greater than 90%.

Acknowledgements

I would first like to thank my advisor, Dr. Anant Madabhushi, for inspiring me with his strong work ethic and indefatigable passion for research. Similarly, my spirited discussions with Dr. Shridar Ganesan, Dr. Michael Feldman, and Dr. John Tomaszewski have yielded valuable insight into the clinical motivation behind my research as well as the challenges faced in digital pathology today.

This work would not have been possible without support from the members of the Laboratory for Computational Imaging and Bioinformatics. Their enthusiasm for research and committment to helping each other succeed has fostered an ideal research environment. I would specifically like to thank Dr. James Monaco, Scott Doyle, Shannon Agner, and George Lee for all of their contributions to this work.

I would be remiss to conclude without mentioning my family. While their contributions to this work are not immediately apparent, my parents Jayanthi and Nagesh Basavanhally and my brother Naveen have been the single largest driving force behind all of my research accomplishments.

This work was made possible by grants from the Wallace H. Coulter Foundation, New Jersey Commission on Cancer Research, National Cancer Institute (Grant Nos. R01CA136535-01,ARRA-NCl-3 R21CA127186-02S1, R21CA127186-01, R03CA128081-01, and R03CA143991-01), The Cancer Institute of New Jersey, and the Life Science Commercialization Award from Rutgers University.

Dedication

This work is dedicated to those responsible for making me who I am today – my parents.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Clinical Motivation

Breast cancer (BC) is the second leading cause of cancer-related deaths in women, with more than 192,000 new cases of invasive BC predicted in the United States for 2009 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), which may have a variety of prognostic implications for BC patients [3, 4, 5, 6]. The function of LI as a potential anti-tumor mechanism in BC was first shown by Aaltomaa et al. [3]. In [4], it was shown that the presence of LI after treatment may predict a patient's response to therapy and long-term disease outcome. Further, Rody et al. [6] suggested that LI extent has the potential to discriminate between estrogen receptor-negative tumors that have good and poor prognoses.

Recently, Alexe et al. [5] demonstrated a correlation between the presence of LI and improved distant recurrence-free survival rates in invasive BC tumors that exhibit amplification of the HER2 gene (HER2+ BC). Since most HER2+ BC is currently treated with agents that specifically target the HER2 protein, the characterization of LI extent may aid clinicians in identifying patients with poor prognoses who may require adjuvant treatments such as chemotherapy. 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 that is not feasible in the clinical setting. The availability of a computerized image analysis system for automated quantification of LI extent in HER2+ BC will enable development of an inexpensive image-based system for predicting disease survival and outcome.

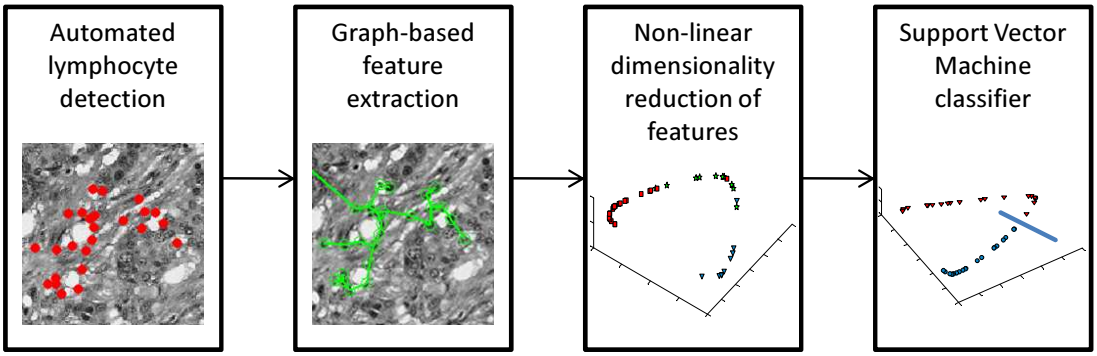

Figure 1.1: A flowchart illustrating the 4 main steps in the CADx scheme. Automated lymphocyte detection is followed by extraction of architectural features. The high-dimensional feature space is then non-linearly embedded into a reduced dimensional space via Graph Embedding, which allows for data visualization and subsequent evaluation via a SVM classifier.

1.2 Brief Outline and Novel Contributions of the Work

In this work, a fully automated CADx system for detecting and grading LI on digitized, hematoxylin and eosin (H & E) stained histopathology is presented. The main components of the CADx methodology is illustrated in the flowchart in Figure 1.1. The methodology comprises three specific aims;

- Aim 1: Automatically detect LI on digitized histopathology images,
- Aim 2: Extract quantitative image-based features to describe LI extent, and
- Aim 3: Visualize the data in a reduced dimensional space based on their LI extent.

Aim 1 is accomplished through an automated detection system that combines a region-growing algorithm with Maximum *a Posteriori* (MAP) estimation and Markov Random Field (MRF) theory [7]. First, all candidate BC and lymphocyte nuclei are detected via a region-growing algorithm that uses contrast measures to find optimal boundaries [8, 7]. By growing outward from the center of each nucleus, this technique is robust to artifacts outside of the nuclei (Figures 2.1(d)-(f)). The region-growing algorithm has a high detection sensitivity, resulting in a large number of lymphocyte and non-lymphocyte 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 2 main classes of objects detected). MRF theory [9, 7] 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 (Figure 2.1(a)) and (2) isolate infiltrating lymphocytes from the surrounding baseline level of lymphocytes. MAP estimation is achieved by using the Iterated Conditional Modes algorithm, a fast and simple method for maximizing the posterior probability that a detected object is indeed a lymphocyte [10, 7].

Detection of LI alone, however, cannot completely characterize the abnormal LI phenotype because a baseline level of lymphocytes is present in all tissues. Hence, quantitative

features must be defined to describe the spatial arrangement of the lymphocyte nuclei. In Aim 2, the centers of individual detected lymphocytes are used as vertices to construct a series of graphs (Voronoi Diagram, Delaunay Triangulation, and Minimum Spanning Tree) and a variety of quantitative signatures describing the spatial arrangement of lymphocytes are extracted from each image [11]. The 50 features used in this CADx system are outlined in Table 3.2. 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, 11].

In Aim 3, a non-linear dimensionality reduction scheme, Graph Embedding [13], is used to project the high-dimensional feature vector into a reduced 3D embedding space for visualization. While a large set of descriptive features is certainly desirable for modeling biological processes such as LI, a high-dimensional feature space also presents two main problems for data classification analysis: (1) the curse of dimensionality [14] affects computational efficiency due to the exponential increase in data volume required for each additional feature and (2) it is impossible to directly visualize the relationships between images in a high-dimensional space. Both of these issues are addressed by dimensionality reduction (DR), which refers to a class of techniques that reduce high-dimensional data into a low-dimensional subspace while preserving object-class relationships from the high-dimensional space [15]. Thus two objects that are close to one another in the original feature space will be mapped to adjacent locations in the low-dimensional subspace. The projection of the image-derived features into a reduced dimensional space via Graph Embedding allowed for the visualization of a smooth manifold which revealed a continuum between low, intermediate, and high levels of LI.

1.3 Organization of the Thesis

The rest of the thesis work is organized as follows. In Chapter 2 previous related work in CADx for BC histopathology is discussed. The details of the methodology are explained in Chapter 3 and a variety of methods for evaluating the efficacy of the CADx system are outlined in Chapter 4. In Chapter 5 quantitative and qualitative results from a cohort of 41 H & E stained, HER2+ BC histopathology images are presented. Concluding remarks and directions for future research are presented in Chapter 6.

Chapter 2

Related Work in Computer-Aided Diagnosis for Digitized Histopathology

Although there has been no significant research into the automated and computerized characterization of LI in H & E stained histopathology imagery, a number of computer-aided diagnosis (CADx) tools have previously been developed for the analysis of BC histopathology.

CADx refers to the quantitative, computerized analysis of biomedical data to assist a physician in characterizing a patient's malignancy. The majority of automated image analysis systems for BC have been limited to computer-aided detection techniques for radiological studies [16, 17]. More recently, researchers [18, 19, 20, 21, 22, 23, 11] have begun to develop computer-aided diagnosis (CADx) schemes for the analysis of digitized BC histopathology; however, their work has mostly focused on either finding suspicious regions of interest (ROI) [18, 21] or has attempted to determine cancer grade from manually isolated ROIs [19, 20, 11]. The methods for both applications use image-based features to discriminate between 2 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 [18, 11]. Textural features and filter banks have also been employed [18, 19, 20, 21, 23] to model the phenotypic appearance of BC histopathology.

2.1 Automated Nuclear Detection

An important prerequisite for extracting histopathological image attributes to model BC appearance is the ability to automatically detect and segment histological structures. Consequently the ability of an image analysis system to grade the extent of LI in a BC histopathology image is dependent on the algorithm's ability to automatically detect lymphocytes. Automated LI detection, however, is a non-trivial task complicated by the intrinsic similarity in appearance of BC nuclei and lymphocyte nuclei on H & E stained breast biopsy samples (Figure 2.1(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 (Figures 2.1(b), (c)) [24]. Biological differences such as the presence of fat deposits (Figure 2.1(d)) can confound algorithms that rely on boundary detection alone. Preparation issues such as "cutting artifact" (Figure 2.1(e)) and digitization misalignment (Figure 2.1(f)) lead to similar problems, but are more difficult to predict and correct for since they are unrelated to the underlying biology.

While several researchers have been developing algorithms for detection of nuclei [25, 26, 27, 20, 28, 29, 30, 31] in digitized histopathology, there have been no significant attempts to automatically detect LI in BC histopathology. Some popular approaches to automated nuclear detection are based on adaptive thresholding [25, 20] and fuzzy c-means clustering [27, 29]. 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 (Figure 2.1(a)). Techniques such as active contours [26, 30, 31] 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 (Figures 2.1(b),

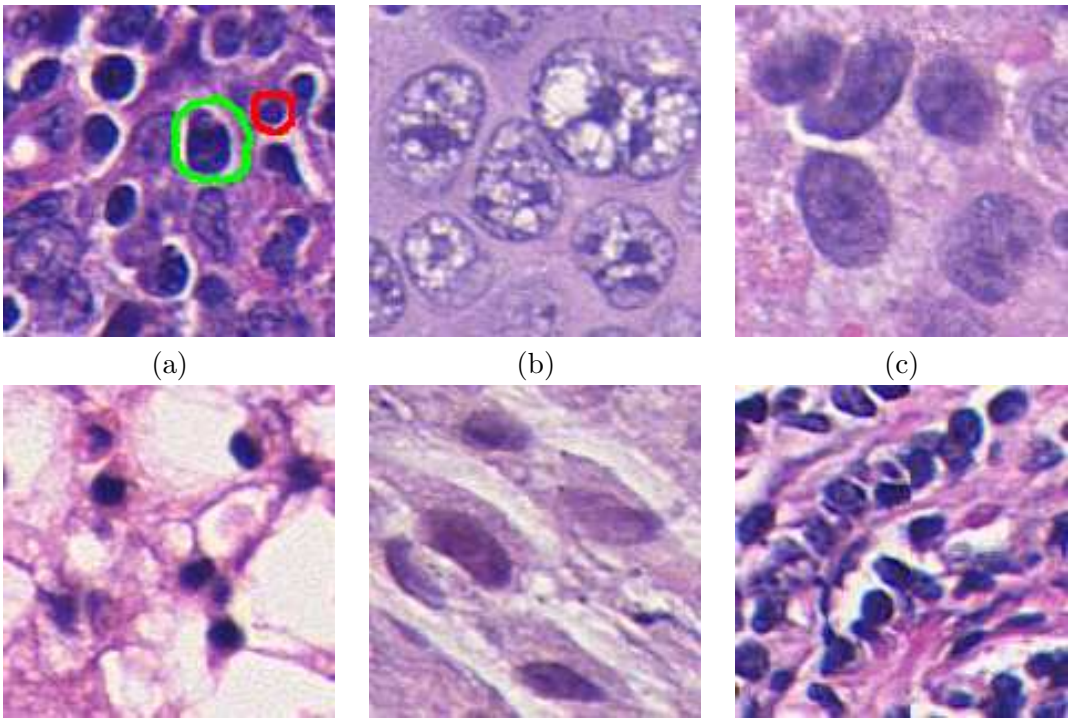

Figure 2.1: There are several challenges in automated LI detection including (a) the similarity in appearance between a cancer cell nucleus (circled in green) and a lymphocyte nucleus (circled in red). In general, lymphocyte nuclei are distinguished from cancer cell nuclei by their smaller size, more circular shape, and a darker, homogeneous staining. Additional challenges include variations in the appearance of (b), (c) BC nuclei within a single histopathology slide, (d) the presence of fat among cancerous tissue, (e) histological fixing, and (f) slide digitization artifacts.

(c)) and image acquisition artifacts (Figures 2.1(e), (f)). Some researchers have developed hybrid techniques in order to improve nuclear detection and segmentation results. For example, Glotsos et al. [30] used Support Vector Machine clustering to improve initialization for active contour models. More recently, semi-automated probabilistic models have used pixel-wise intensity information to detect cancer [28] and lymphocyte nuclei [32] in digitized BC histopathology. Probabilistic models, however, are usually limited by the availability of expert-annotated training data.

2.2 Feature Extraction

The need to quantify the morphology and arrangement of histological structures has been addressed in previous work. In [33], Sudbo et al. constructed graphs to model tissue architecture in oral mucosa, whereby a graph is defined as a set of vertices (nuclei) with corresponding edges connecting all nuclei. Similarly, Gunduz et al. [22] explored automated cancer diagnosis in brain cancer by using hierarchical graphs. Doyle et al. [12, 11] have previously shown the importance of using graphs to describe the spatial arrangement of nuclei in distinguishing cancer grade in both prostate cancer and BC histopathology. In [11], 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.

Textural features have also been applied successfully to specific applications in computerized histopathology analysis. Traditional methods such as first order gray-level, second-order Haralick statistics [34], and Gabor filter features have been used to identify cancerous regions in both breast [11] and prostate cancer [12] histopathology. Other textural features based on wavelets have been used in the computerized analysis of BC histopathology [19]. Additionally, texton signatures [35] have been used successfully in applications related to content-based image retrieval [36] and computer-aided classification [37] of digitized cancer histopathology.

2.3 Non-linear Dimensionality Reduction

The digital pathology field is known to generate large amounts of data; hence, dimensionality reduction (DR) methods can be used to distill the most relevant components of the data

and make further analysis more feasible. Linear DR methods such as principal component analysis assume linear relationships between all data in the high-dimensional space. Since it has previously been demonstrated that biomedical data is non-linear in nature [15], the use of linear DR methods may produce suboptimal results. Conversely, non-linear DR methods such as Graph Embedding [13], Locally Linear Embedding [38], and Isometric Mapping [39] embed the data into a low-dimensional embedding while attempting to preserve the global structure of the data manifold by maintaining geodesic distances between objects. The role of non-linear dimensionality reduction as a visualization tool for the analysis of digitized histopathology has previously been demonstrated by Doyle et al. in both prostate cancer grading [12] and breast cancer grading [11] tasks.

Chapter 3

Methodology

3.1 Notation

A dataset $\mathbf{Z} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_\mathcal{M}\}$ is defined as a set of $\mathcal{M}$ images. An image scene $\mathcal{C} \in \mathbf{Z}$ is defined as $\mathcal{C} = (C, g)$, where C is a 2D set of pixels $c \in C$ and g is the associated luminance function from the CIE-Lab color space [40]. The CIE-Lab color space is used in this work because it has the advantage of being more perceptually uniform and more robust to variations in staining and digitization than RGB space (Figure 2.1(a)) [40, 21]. A list of symbols and notation commonly used in this thesis is shown in Table 3.1.

3.2 Aim 1: Automated Detection of Lymphocytic Infiltration

Beginning with a set of N candidate lymphocyte nuclear centers $\mathbf{N} = \{n_1, n_2, \ldots, n_N\}$, a set of L finalized lymphocyte nuclei is identified with centers given by $\mathcal{O} = \{o_1, o_2, \ldots, o_L\}$, such that $\mathcal{O} \subseteq \mathbf{N}$. The following sections detail the region-growing, Maximum *a Posteriori* (MAP) estimation, and Markov Random Field (MRF) algorithms that comprise the lymphocyte detection module of the CADx system (Figure 3.1).

Symbol	Description
$\mathbf{Z} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_\mathcal{M}\}$	HER2+ BC histopathology dataset comprising $\mathcal{M}$ digitized images
$\mathcal{C} = (C, g)$	Image scene defined by a set of pixels (C) and luminance function (g)
$\mathbf{N} = \{n_1, n_2, \ldots, n_N\}$	N candidate lymphocyte nuclei centers in image scene $\mathcal{C}$
$\mathbf{R} = \{r_1, r_2, \ldots, r_N\}$	N candidate regions grown from $\mathbf{N}$
$\mathcal{O} = \{o_1, o_2, \ldots, o_L\}$	L finalized lymphocyte nuclei centers in image scene $\mathcal{C}$, where $\mathcal{O} \subseteq \mathbf{N}$
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 $r \in \mathbf{R}$
$Y_r = [A_r, \sigma_r]^\mathsf{T} \in \mathbb{R}^{+2}$	Random variable denoting features square root of area (A) and std. dev. intensity (σ) for each region $r \in \mathbf{R}$
x_r, y_r	Specific instances of X_r and Y_r
$\mathbf{x}, \mathbf{y}$	Sets of $x_r, \forall r \in \mathbf{R}$ and $y_r, \forall r \in \mathbf{R}$
$\mathcal{G} = \{\mathcal{O}, \mathbf{E}, \mathbf{W}\}$	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.

Table 3.1: A list of key notation used in this thesis.

3.2.1 Candidate Lymphocyte Detection via Region-Growing

First, candidate image locations that could represent centers of lymphocytic nuclei are identified. 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 (Figure 2.1). The image scene $\mathcal{C}$ is convolved with a Gaussian (smoothing) kernel at multiple scales $\sigma_G \in \{6, 7, 8\}$ µm 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 $\mathbf{N} = \{n_1, n_2, \ldots, n_N\}$ that represent candidate lymphocyte centers on the original scene $\mathcal{C}$. Each $n \in \mathbf{N}$ is grown into a corresponding region $r \in \mathbf{R}$ using the

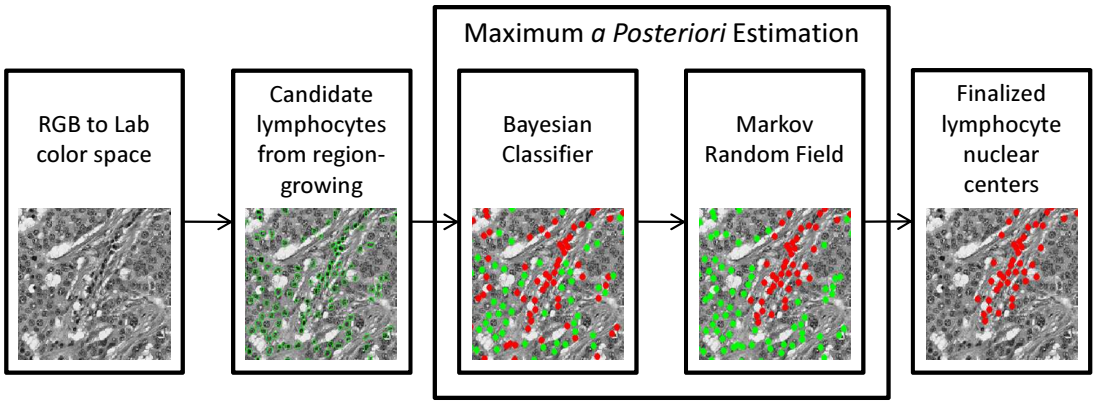

Figure 3.1: A flowchart illustrating the main steps in the automated lymphocyte detection scheme.

4-step procedure described below. Additional details on the region-growing algorithm can be found in [8].

Step 1: A set of current pixels $S_{\mathrm{CR}} = \{n\}$ is initialized as shown in Figure 3.2(a). The current boundary S_{CB} is defined as the set of 8-connected pixels surrounding S_{CR}. A square bounding box S_{BB} containing all pixels within a $12\sigma_G \times 12\sigma_G$ neighborhood around n is then constructed.

Step 2: The pixel $c \in S_{\mathrm{CB}}$ 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 $d \in S_{\mathrm{BB}}$ that is an 8-connected neighbor of c and $d \notin S_{\mathrm{CR}}$. A set of internal boundary pixels $S_{\mathrm{IB}} \subset S_{\mathrm{CR}}$ (Figures 3.2(b), (c)) is defined as all pixels in S_{CR} that are 8-connected to any pixel in S_{CB}.

Step 3: $\overline{g}_{\mathrm{IB}}$ and $\overline{g}_{\mathrm{CB}}$ are computed as the mean intensity of pixels in S_{IB} and S_{CB}, respectively. The boundary strength is computed at each iteration as $\overline{g}_{\mathrm{IB}} - \overline{g}_{\mathrm{CB}}$.

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 $\overline{g}_{\mathrm{IB}} - \overline{g}_{\mathrm{CB}}$ is maximum (Figures 3.4(b), (f)).

Since the region-growing procedure is repeated with seed points from a variety of smoothing scales $\sigma_G \in \{6, 7, 8\}$ µm, overlapping regions are resolved by discarding the region with the lower boundary strength. For the sake of convenience the symbol $R \equiv S^*_{\mathrm{CR}}$ will be used through the rest of this thesis.

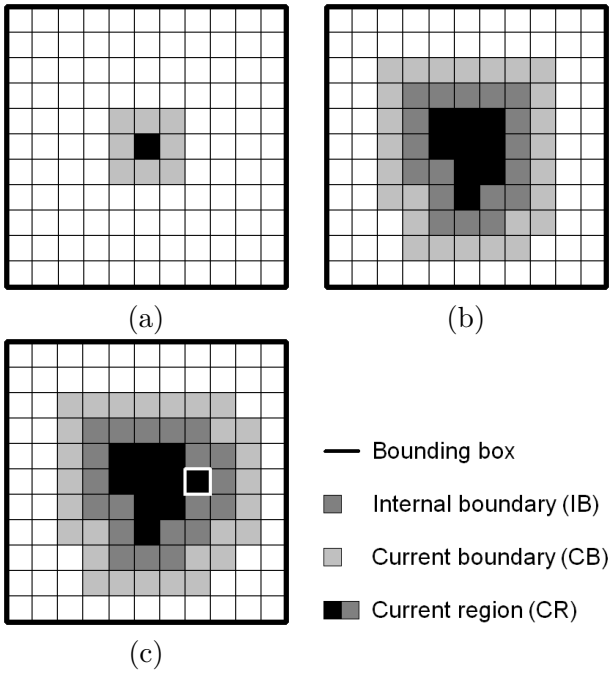

Figure 3.2: Schematic illustrating the iterative growth of a region r. After initialization of the current region S_{CR} (Figure 3.2(a)), current boundary S_{CB}, and bounding box S_{BB}, new pixels are added iteratively (Figure 3.2(b)). When a new pixel (outlined in white) is added to S_{CR}, the boundaries S_{CB} and S_{IB} are adjusted accordingly (Figure 3.2(c)).

3.2.2 Bayesian Modeling of Lymphocytic Infiltration via Maximum a Posteriori 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 breast cancer (ω_b) or

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

$$A_r = \sqrt{|R|}, \tag{3.1}$$

$$\sigma_r = \sqrt{\frac{1}{|R|}\sum_{c\in R}(g(c) - \overline{g})^2}, \tag{3.2}$$

where A_r is the square root of nuclear area (Equation 3.1), σ_r is the standard deviation of luminance in the nuclear region (Equation 3.2), $|R|$ is the cardinality of R, and $\overline{g} = \frac{1}{|R|}\sum_{c\in R} g(c)$ is the average pixel intensity of R. The choice of the two features (A_r and σ_r) is motivated by the fact that (a) BC nuclei are typically larger than lymphocyte nuclei and (b) 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 $y_r = [A_r, \sigma_r]^\mathsf{T} \in \mathbb{R}^{+2}$, respectively. The random variables are defined collectively for all $r \in \mathbf{R}$ as $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$ and $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_N\}$ with state spaces $\Omega = \Lambda^N$ and $\mathbb{R}^{+2\times N}$, respectively. Instances of $\mathbf{X}$ and $\mathbf{Y}$ are denoted by variables $\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \Omega$ and $\mathbf{y} = (y_1, y_2, \ldots, y_N) \in \mathbb{R}^{+2\times N}$.

The labels $\mathbf{X} = \mathbf{x}$, given the feature vectors $\mathbf{Y} = \mathbf{y}$, are estimated using Maximum *a Posteriori* (MAP) estimation [41], which advocates finding the $\mathbf{x}$ that maximizes the posterior probability

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}), \tag{3.3}$$

where $p(\mathbf{y}|\mathbf{x})$ is the likelihood term and $p(\mathbf{x}), p(\mathbf{y})$ are prior distributions for $\mathbf{x}$ and $\mathbf{y}$ respectively. Since maximization of Equation 3.3 is only with respect to $\mathbf{x}$, the prior distribution $p(\mathbf{y})$ is ignored.

Modeling Lymphocyte Features via Trained Probability Distributions

The likelihood term $p(\mathbf{y}|\mathbf{x})$ in Equation 3.3 is calculated from probability density functions (PDFs), where $\mathbf{x}$ is provided by manual delineation of lymphocytes in a training set. Under

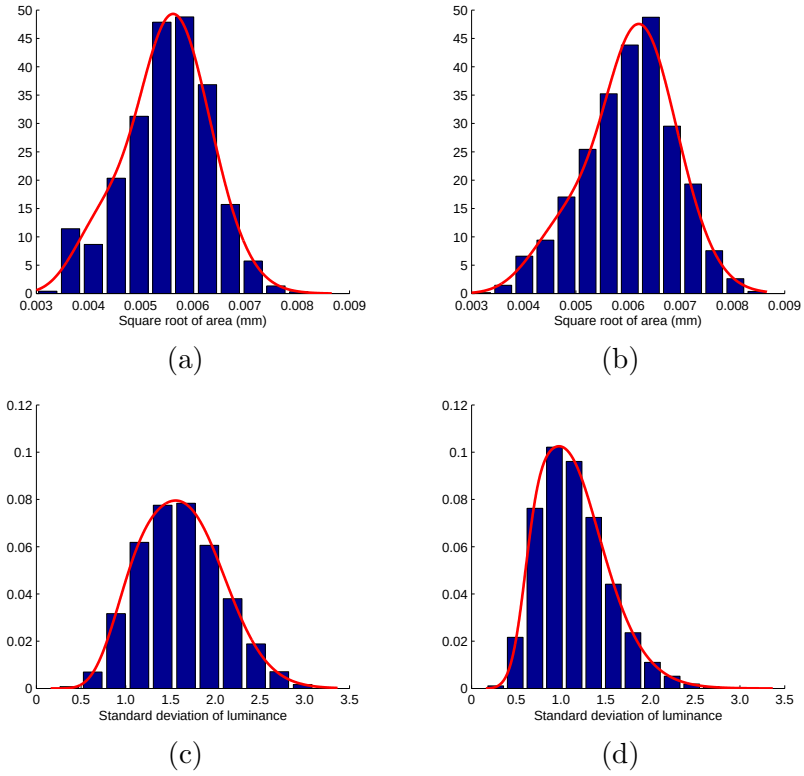

Figure 3.3: Probability density functions (PDFs) estimated from empirical training data and modeled via weighted sum of gamma distributions for (a), (c) ω_ℓ and (b), (d) ω_b classes for (a), (b) square root of area and (c), (d) variance in luminance of each $r \in \mathbf{R}$. In each distribution (a)-(d), the estimated parametric model is overlaid.

the assumption that $\mathbf{y}$ is independent and identically distributed, the likelihood term in Equation 3.3 can be simplified such that

$$p(\mathbf{y}|\mathbf{x}) = \prod_{r \in \mathbf{R}} p(y_r|x_r). \tag{3.4}$$

Each 2-dimensional probability density function (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 four one-dimensional PDFs $\mathcal{F}(A_r|\omega_b)$, $\mathcal{F}(A_r|\omega_\ell)$, $\mathcal{F}(\sigma_r|\omega_b)$, and $\mathcal{F}(\sigma_r|\omega_\ell)$ are required as shown in Figure 3.3. To reduce local irregularities and create a smooth, continuous distribution, the one-dimensional PDFs

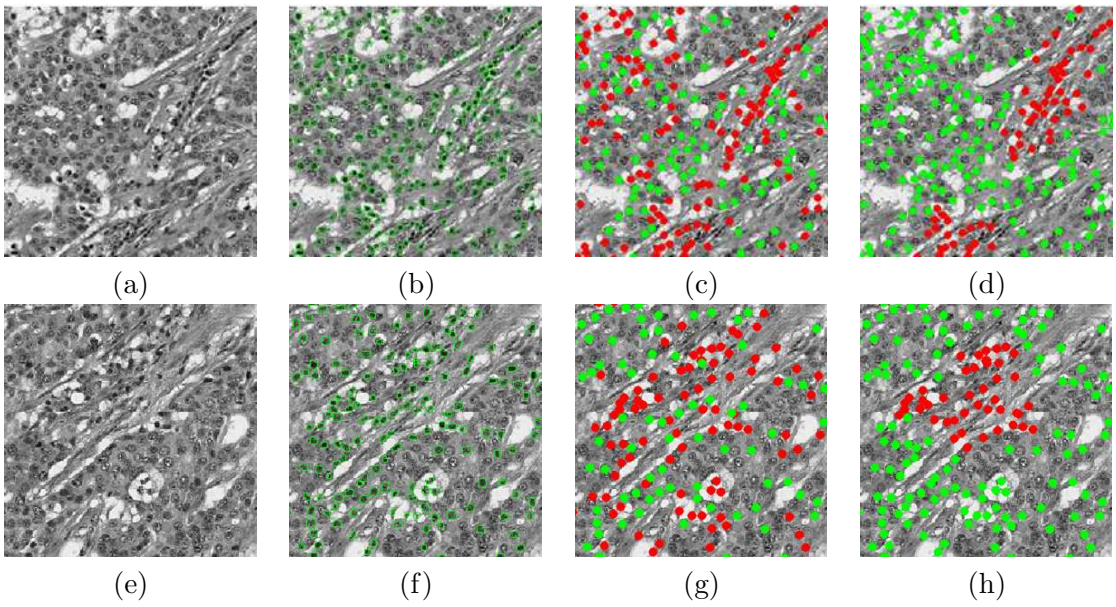

Figure 3.4: (a), (e) Luminance channels of two different HER2+ BC histopathology studies and corresponding results for (b), (f) initial region-growing based lymphocyte detection, (c), (g) preliminary Bayesian refinement showing detected BC nuclei in green and detected lymphocyte nuclei in red, and (d), (h) final lymphocyte detection result after the MRF pruning step.

are modeled by mixtures of Gamma distributions [42]

$$\overline{\Gamma}(z;\delta,\boldsymbol{\phi},\boldsymbol{t}) = \delta z^{t_1-1}\frac{e^{-z/\phi_1}}{\phi_1^{t_1}\Gamma(t_1)} + (1-\delta)z^{t_2-1}\frac{e^{-z/\phi_2}}{\phi_2^{t_2}\Gamma(t_2)}, \tag{3.5}$$

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 [42]. Thus Equation 3.3 can be estimated by calculating $p(\mathbf{y}|\mathbf{x})$ and tentative classes $x_r \in \{\omega_b, \omega_\ell\}$ can be assigned to each $r \in \mathbf{R}$ (Figures 3.4(c), (g)).

Modeling Lymphocyte Proximity via Markov Random Fields

The prior distribution $p(\mathbf{x})$ (Equation 3.3) is defined by a Markov Random Field (MRF). The Markov property [9] states that

$$p(x_r|\mathbf{x}_{-r}) = p(x_r|\mathbf{x}_{\eta_r}), \tag{3.6}$$

where the neighborhood η_r is empirically assumed to contain all regions within a 30 µm radius of r, $\mathbf{x}_{-r} = \{x_s : s \in \mathbf{R}, s \neq r\}$, and $\mathbf{x}_{\eta_r} = \{x_s : s \in \eta_r\}$. The Iterated Conditional Modes (ICM) algorithm [10], a deterministic relaxation procedure, is used to perform MAP estimation (Equation 3.3) and assign a hard label $x_r \in \{\omega_b, \omega_\ell\}$ to each $r \in \mathbf{R}$. Thus each object is classified as either a BC or lymphocyte nucleus (Figures 3.4(d), (h)). The objects labeled as BC nuclei are discarded, while centers of the L lymphocyte nuclei regions are computed and stored as $\mathcal{O} = \{o_1, o_2, \ldots, o_L\}$.

3.3 Aim 2: Architectural Feature Extraction

The complete, undirected graph $\mathcal{G} = (\mathcal{O}, \mathbf{E}, \mathbf{W})$, where $\mathcal{O} = \{o_1, o_2, \ldots, o_L\}$ is the set of vertices corresponding to the set of lymphocyte nuclear centroids, $\mathbf{E} = \{E_1, E_2, \ldots, E_m\}$ is the set of edges connecting the nuclear centroids such that $\{(o_i, o_j) \in \mathbf{E} : \forall o_i, o_j \in \mathcal{O},\ i, j \in \{1, 2, \ldots, L\},\ i \neq j\}$, and $\mathbf{W} = \{W_1, W_2, \ldots, W_m\}$ is a set of weights proportional to the length of each $E \in \mathbf{E}$. To extract information about the arrangement of lymphocyte nuclei, subgraphs are constructed representing the Voronoi Diagram $\mathcal{G}_\mathcal{V}$, Delaunay Triangulation $\mathcal{G}_\mathcal{D}$, and Minimum Spanning Tree $\mathcal{G}_{\text{MST}}$ (Figure 3.5). In addition, statistics describing the number and density of nuclei are calculated directly from $\mathcal{O}$.

3.3.1 Voronoi Diagram

The Voronoi graph $\mathcal{G}_\mathcal{V} = (\mathcal{O}, \mathbf{E}_\mathcal{V}, \mathbf{W}_\mathcal{V})$ (Figures 3.5(b), (f)) is a spanning subgraph of $\mathcal{G}$ defined as a set of polygons $\mathbf{P} = \{P_1, P_2, \ldots, P_L\}$ surrounding all nuclear centroids $\mathcal{O}$ [33, 11]. 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 ($\mathbf{f}_{\mathcal{V}}$) for each scene $\mathcal{C}$ (Table 3.2).

3.3.2 Delaunay Triangulation

The Delaunay graph $\mathcal{G}_{\mathcal{D}} = (\mathcal{O}, \mathbf{E}_{\mathcal{D}}, \mathbf{W}_{\mathcal{D}})$ (Figures 3.5(c), (g)) is a spanning subgraph of $\mathcal{G}$ and the dual graph of $\mathcal{G}_{\mathcal{V}}$ [11]. It is constructed such that if $P_i, P_j \in \mathbf{P}$ share a side, where $i, j \in \{1, 2, \ldots, L\}$, their nuclear centroids $o_i, o_j \in \mathcal{O}$ are connected by an edge $(o_i, o_j) \in \mathbf{E}_{\mathcal{D}}$. The mean, standard deviation, min/max ratio, and disorder are calculated for the side length and area of all triangles in $\mathcal{G}_{\mathcal{D}}$, yielding a set of 8 features ($\mathbf{f}_{\mathcal{D}}$) for each scene $\mathcal{C}$ (Table 3.2).

3.3.3 Minimum Spanning Tree

A spanning tree $\mathcal{G}_{\mathcal{S}} = (\mathcal{O}, \mathbf{E}_{\mathcal{S}}, \mathbf{W}_{\mathcal{S}})$ refers to any spanning subgraph of $\mathcal{G}$ [11]. The total weight $\widehat{\mathbf{W}}_{\mathcal{S}}$ for each subgraph is determined by summing all individual weights $W \in \mathbf{W}_{\mathcal{S}}$. The Minimum Spanning Tree $\mathcal{G}_{\text{MST}}$ (Figures 3.5(d), (h)) is the spanning tree with the lowest total weight such that $\mathcal{G}_{\text{MST}} = \arg\min_{\mathcal{G}_{\mathcal{S}} \in \mathcal{G}} \left[\widehat{\mathbf{W}}_{\mathcal{S}}\right]$. The mean, standard deviation, min/max ratio, and disorder of the branch lengths in $\mathcal{G}_{\text{MST}}$ yield a set of 4 features ($\mathbf{f}_{\text{MST}}$) for each scene $\mathcal{C}$ (Table 3.2).

3.3.4 Nuclear Features

The global density $\frac{L}{|C|}$ 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 $o_i \in \mathcal{O}$, a corresponding nuclear neighborhood $\eta^{\zeta}(o_i) = \{o_j : \|o_i - o_j\|_2 < \zeta, o_j \in \mathcal{O}, o_j \neq o_i\}$ is defined, where $\zeta \in \{10, 20, \ldots, 50\}$ 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 the minimum radius ζ^* is found such that $|\eta^{\zeta^*}(o_i)| \in \{3, 5, 7\}$ and the mean, standard deviation, and disorder are calculated over all $o_i \in \mathcal{O}$. A total of 25 nuclear features ($\mathbf{f}_{\text{NF}}$) are extracted for each scene $\mathcal{C}$ (Table 3.2).

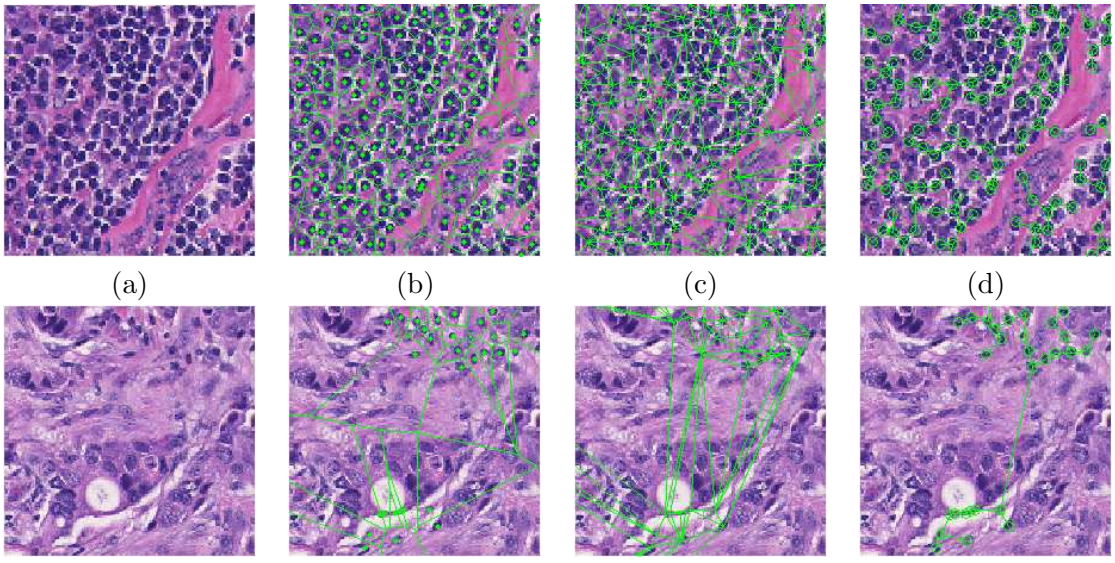

Figure 3.5: Two different HER2+ breast cancer histopathology images with (a) high and (b) low levels of LI. Figures 3.5((b), (f)) show the corresponding Voronoi Diagrams constructed using the automatically detected lymphocyte centers as vertices of the graph. Corresponding Delaunay Triangulation and Minimum Spanning Tree graphs are shown in Figures 3.5((c), (g)) and 3.5((d), (h)), respectively.

3.4 Aim 3: Non-linear Dimensionality Reduction via Graph Embedding

Graph Embedding (GE) is employed to non-linearly transform the high-dimensional set of image features into a low-dimensional embedding while preserving relative distances between images from the original feature space [11]. For each scene $\mathcal{C}$, a 50-dimensional image feature set is defined as the superset $\mathbf{F} = \{\mathbf{f}_{\mathcal{V}}, \mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\text{MST}}, \mathbf{f}_{\text{NF}}\}$ containing all features derived from the Voronoi Diagram, Delaunay Triangulation, Minimum Spanning Tree, and

Feature Set	Description	No. of features
$\mathbf{f}_{\mathcal{V}}$	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	
$\mathbf{f}_{\mathcal{D}}$	Triangle side length: mean, std dev., min/max ratio, disorder	8
	Triangle area: mean, std dev., min/max ratio, disorder	
$\mathbf{f}_{\text{MST}}$	Edge length: mean, std dev., min/max ratio, disorder	4
$\mathbf{f}_{\text{NF}}$	Density of nuclei	25
	Distance to $\{3, 5, 7\}$ nearest nuclei: mean, std dev., disorder	
	Nuclei in $\zeta \in \{10, 20, \ldots, 50\}$ pixel radius: mean, std dev., disorder	

Table 3.2: A breakdown of the 50 architectural features, comprising 25 graph-based and 25 nuclear attributes.

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 $a, b \in \{1, 2, \ldots, \mathcal{M}\}$, 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,

$$\mathcal{E}(\mathbf{F}') = 2(\mathcal{M} - 1) \cdot \text{trace}\left[\frac{\mathbf{F}'^{\top}(\mathcal{A} - \mathcal{W}_{\mathbf{F}})\mathbf{F}'}{\mathbf{F}'^{\top}\mathcal{A}\mathbf{F}'}\right], \tag{3.7}$$

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

Chapter 4

Evaluation Methods

4.1 Quantitative Evaluation of Automated LI Detection via Hausdorff Distance

The automated lymphocyte detection algorithm is evaluated by the Hausdorff distance [43], 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 $\mathcal{O}^{\text{man}}$ were estimated exhaustively by an expert pathologist who manually annotated the individual lymphocyte nuclei in each scene. The partial, directed Hausdorff distance is calculated for $\mathcal{O}^{\text{auto}}$ with respect to $\mathcal{O}^{\text{man}}$ as,

$$\Phi_H(\mathcal{O}^{\text{auto}}, \mathcal{O}^{\text{man}}) = \min_{u \in \mathcal{O}^{\text{man}}} \|v - u\|_2, \forall v \in \mathcal{O}^{\text{auto}}. \tag{4.1}$$

4.2 Quantitative Evaluation of Architectural Features via Support Vector Machine Classifier

The SVM classifier [44] is employed to evaluate the ability of the image descriptors to discriminate between high and low levels of LI in histopathology images. The SVM classifier is constructed by using a Gaussian kernel function 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 $\mathbf{F}'(\mathcal{C}_b)$, $a, b \in \{1, 2, \ldots, \mathcal{M}\}$, respectively, the Gaussian kernel $\Pi(\mathbf{F}'(\mathcal{C}_a), \mathbf{F}'(\mathcal{C}_b)) = \exp(-\epsilon\left(\|\mathbf{F}'(\mathcal{C}_a) - \mathbf{F}'(\mathcal{C}_b)\|_2\right)^2)$, 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 [28]. The general form of the SVM is given as,

$$\Theta(\mathcal{C}_a) = \sum_{\gamma=1}^{\tau} \xi_\gamma \mathcal{Y}(\mathcal{C}_\gamma) \Pi(\mathbf{F}'(\mathcal{C}_a), \mathbf{F}'(\mathcal{C}_\gamma)) + \mathbf{b}, \tag{4.2}$$

where $\gamma \in \{1, 2, \ldots, \tau\}$ represents the τ marginal training samples (i.e. support vectors), $\mathbf{b}$ is the hyperplane bias estimated for $\mathbf{Z}_{\text{tra}}$, and ξ_γ is the model parameter determined by maximizing the objective function [44, 15]. 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 $\mathcal{Y}(\mathcal{C}_a) = \text{sign}\left[\Theta(\mathcal{C}_a)\right]$.

The Gaussian kernel has recently become popular for classification in a number of biomedical image processing applications [28, 45]. This CADx algorithm uses the Gaussian kernel instead of the traditional linear kernel [15] because its non-linear projection helps create 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 [41]. A k-fold cross-validation scheme [41] 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. All samples from $k-1$ subsets are pooled together to obtain $\mathbf{Z}_{\text{tra}}$ and the remaining subset is used as $\mathbf{Z}_{\text{tes}}$. For each of the k iterations, an SVM classifier is trained with $\mathbf{Z}_{\text{tra}}$ and evaluated on $\mathbf{Z}_{\text{tes}}$; a new $\mathbf{Z}_{\text{tes}}$ and $\mathbf{Z}_{\text{tra}}$ being chosen at each iteration so that all samples are evaluated. Using a value of $k = 3$, the entire cross-validation algorithm is repeated over 100 trials and the resulting mean (μ_{ACC}) and standard deviation (σ_{ACC}) of the classification accuracy obtained. Classification accuracy is defined as the ratio between the number of correctly classified images and the total number of images in the dataset.

4.3 Formulation of Textural Features

To verify the significance of the architectural features (Section 3.3) to the performance of CADx system, two different sets of texture signatures (Varma-Zisserman textons [35] and global textures [11]) are considered in this work.

4.3.1 Varma-Zisserman Texton-Based Classifier

The Varma-Zisserman (VZ) texton-based features [35] used to distinguish histopathology images with high and low LI extent are calculated as described in the steps below. The reader is referred to [35] for additional details regarding VZ textons.

Step 1: All $\mathcal{C}_{\text{tra}} \in \mathbf{Z}_{\text{tra}}$ are first convolved with the Maximum Response 8 (MR8) filter bank [35], which contains edge and bar filters at several orientations and scales. An 8-dimensional MR8 feature vector $\mathbf{f}_{\text{text}}(c)$ is defined for each $c \in C, \forall \mathcal{C}_{\text{tra}} \in \mathbf{Z}_{\text{tra}}$.

Step 2: Feature vectors $\mathbf{f}_{\text{text}}$ of all $c \in C, \forall \mathcal{C}_{\text{tra}} \in \mathbf{Z}_{\text{tra}}$ are clustered using the K-means algorithm [41] and the K cluster centers $\{c_1^*, c_2^*, \ldots, c_K^*\}$ are defined as textons.

Step 3: For each $c \in \mathcal{C}_{\text{tra}}$, the closest corresponding texton $c_j^*, j \in \{1, 2, \ldots, K\}$ 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 1D 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.

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 $\mathcal{C}_{\text{tes}}$ is identified based on

$$\mathcal{C}_{\text{tra}}^* = \underset{\mathcal{C}_{\text{tra}} \in \mathbf{Z}_{\text{tra}}}{\text{argmin}} \left[\chi^2(\mathcal{H}(\mathcal{C}_{\text{tra}}), \mathcal{H}(\mathcal{C}_{\text{tes}}))\right], \tag{4.3}$$

where $\chi^2(\mathcal{H}(\mathcal{C}_{\text{tra}}), \mathcal{H}(\mathcal{C}_{\text{tes}}))$ is the Chi-squared distance [46] between the histograms of $\mathcal{C}_{\text{tra}}$ and $\mathcal{C}_{\text{tes}}$. If $\mathcal{Y}(\mathcal{C}_{\text{tes}}) = \mathcal{Y}(\mathcal{C}_{tra}^*)$, $\mathcal{C}_{\text{tes}}$ is said to have been correctly classified; otherwise incorrectly classified. Additional details on the VZ texton approach can be found in [35].

The mean μ_{ACC} and standard deviation σ_{ACC} of the classification accuracy of the VZ-texton approach are calculated over 100 randomized 3-fold cross-validation trials (Table 5.1). These experiments are repeated for each $K \in \{2, 3, 5, 10\}$.

4.3.2 Global Texture Features

Three types of global texture features are extracted in this work: (1) first order gray-level features, (2) second order Haralick statistics, and (3) Gabor filter features. In each image, these signatures are calculated for each channel in the HSI (hue, saturation, intensity) color space at three window sizes (3x3, 5x5, and 7x7 pixels) [12, 11].

First order gray-level features: First order gray-level features are calculated directly from the HSI values in each $\mathcal{C}$ [12]. The mean, standard deviation, minimum-to-maximum ratio (min/max), and mode are calculated over all $c \in C$ to yield a total of 540 features for each image scene $\mathcal{C}$ [11].

Second order Haralick statistics: Second-order co-occurrence texture features are described by the 16 Haralick statistics presented in [34]. For each image scene $\mathcal{C}$, a co-occurrence matrix is generated and 16 Haralick scenes are calculated. The average, standard deviation, min/max ratio, and mode of the values in each Haralick scene are calculated to yield 576 second order features for each $\mathcal{C}$ [11].

Gabor filter features: Steerable Gabor filters respond to a variety of textural differences in an image. A unique filter kernel $\mathfrak{G}$ is defined as shown in [12]. A total of 64 Gabor filter responses are generated by varying the orientation parameter over $\{0, \frac{\pi}{8}, \frac{2\pi}{8}, \ldots, \frac{7\pi}{8}\}$ and scale parameter over $\{0, 1, \ldots, 7\}$. The average, standard deviation, min/max ratio, and mode over all $c \in C$ are calculated to yield a total of 2,304 Gabor feature values for each $\mathcal{C}$ [11].

Chapter 5

Results and Discussion

5.1 Dataset

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 20x optical magnification (0.33 µm 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 3 classes by a BC oncologist based on LI extent. The dataset comprises 22 low, 10 medium, and 9 high LI samples. For the purpose of quantitative classification (as described in Section 4.2), the oncologist separated the images into two classes comprising 22 low LI and 19 high LI samples, respectively.

5.2 Performance of Automated LI Detection

Over a total of $|\mathcal{O}^{\text{auto}}| = 42,000$ automatically detected lymphocyte nuclei for all $\mathcal{C} \in \mathbf{Z}$, the median partial Hausdorff distance was determined to be 3.70 µm (Figure 5.1). Considering an average lymphocyte diameter of approximately 7 µm, these results verify the ability of the algorithm 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 5.3 and Table 5.1.

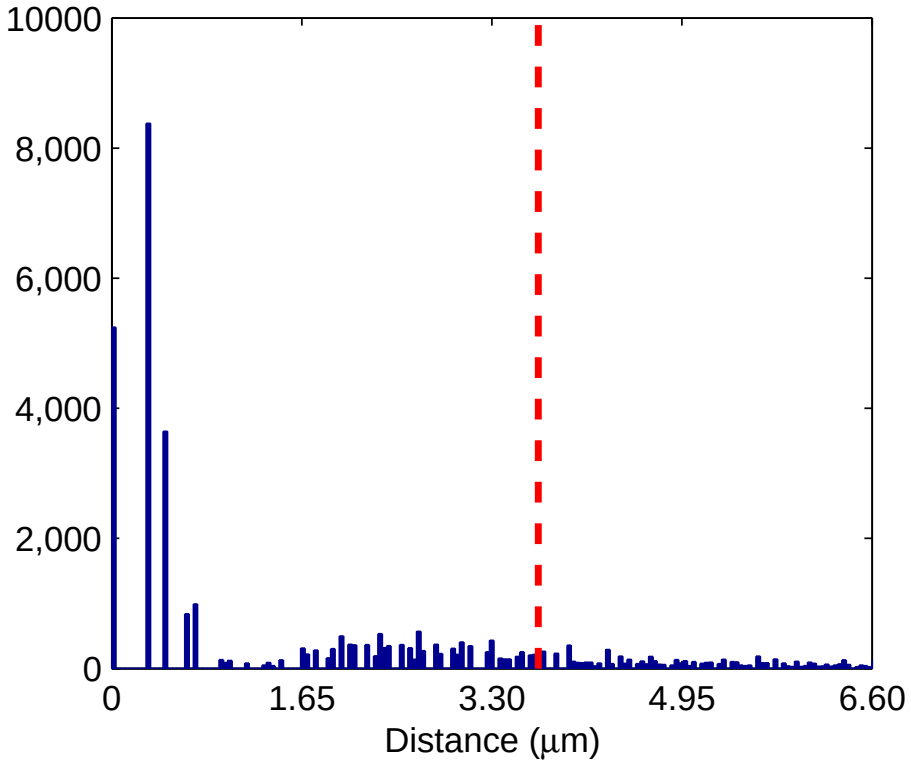

Figure 5.1: A histogram of the partial, directed Hausdorff distances $\Phi_H(\mathcal{O}^{\text{auto}}, \mathcal{O}^{\text{man}})$ between automatically and manually detected lymphocyte nuclei in all 41 HER2+ BC histopathology images. The red dashed line denotes the median of the errors of the automated lymphocyte detection scheme.

5.3 Performance of Architectural Features

Table 5.1 shows the classification accuracies of the 3-dimensional reduced feature set $\mathbf{F}'(\mathbf{Z})$ resulting from both automated and manual LI detection via the SVM classifier. Note that the 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 the LI detection algorithm. Table 5.1 also reveals that the original architectural features $\mathbf{F}(\mathbf{Z})$ (via automated LI 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.

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
Gray-level	50.22 ± 6.22
Haralick	50.88 ± 7.62
Gabor filter	54.22 ± 6.48

Table 5.1: Results of SVM classification accuracy (μ_{ACC}, σ_{ACC}) for 41 BC histopathology images using 100 3-fold cross-validation trials for automated and manual lymphocyte detection with the architectural (both reduced $\mathbf{F}'$ and unreduced $\mathbf{F}$), VZ texton classifier, and global texture features.

In order to determine the optimal dimensionality for performing classification, the architectural feature set $\mathbf{F}(\mathbf{Z})$ was reduced to various dimensionalities $\{2, 3, \ldots, 10\}$ via Graph Embedding. For each dimensionality, the corresponding μ_{ACC} and error bars (σ_{ACC}) over 100 trials of randomized 3-fold cross-validation were calculated (Figure 5.2). Figure 5.2 suggests that classification accuracy is stable at lower dimensionality and drops off slightly at higher dimensionality.

5.4 Performance of Textural Features

The classification results (Table 5.1) show that the Varma-Zisserman (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% ± 5.66%. Similarly, the best performance by the global texture features resulted in a classification accuracy of 54.22% ± 6.48%. These result suggests that texture

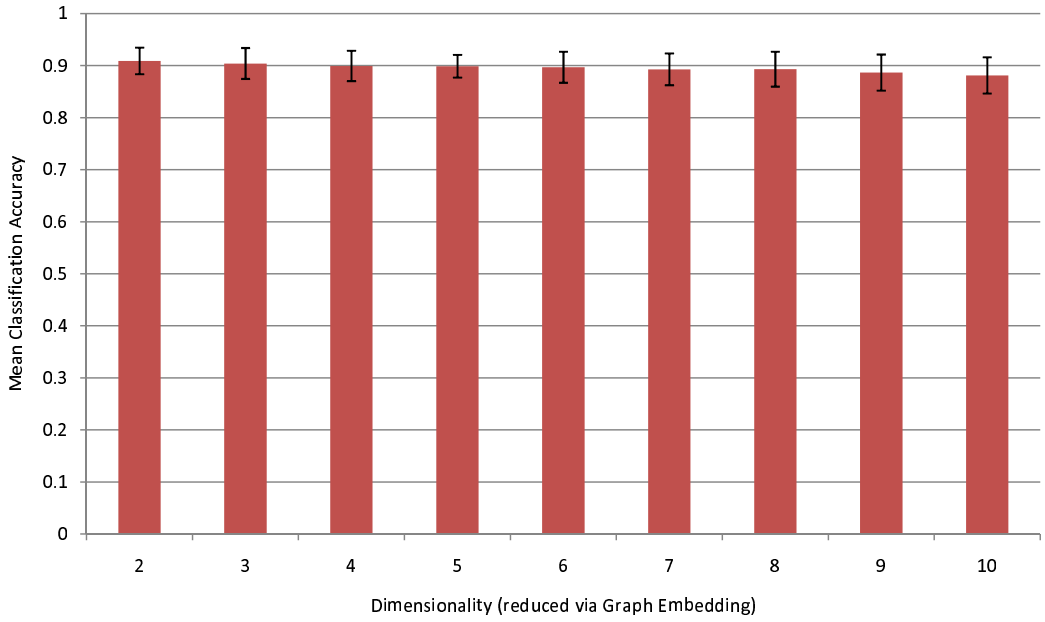

Figure 5.2: The mean (μ_{ACC}) classification accuracy over 100 trials of 3-fold cross-validation is shown for different dimensionalities $\{2, \ldots, 10\}$ obtained via Graph Embedding. The error bars represent standard deviation (σ_{ACC}) of the classification accuracy.

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 (Figure 2.1) 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.

5.5 Low-Dimensional Manifold Visualization

Apart from helping 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. Figure 5.3 shows the reduced dimensional representation ($\beta = 3$ dimensions) of the high dimensional architectural and VZ-texture feature spaces. Note that the 3 axes

in each of Figures 5.3(a)-(d) reflect the principal Eigen vectors obtained embedding the data via GE. Reducing the architectural feature set to 3 dimensions via Graph Embedding reveals the progression from low to medium to high degrees of LI on a smooth, continuous manifold (Figures 5.3(a), (b)). Conversely, the VZ features (Figure 5.3(c)) and global texture features (Figure 5.3(d)) neither produce a continuous manifold, nor appear to stratify samples based on LI extent. The plots in Figure 5.3 further validate the quantitative classification results shown in Table 5.1 and reflect the efficacy of architectural image features in stratifying extent of LI.

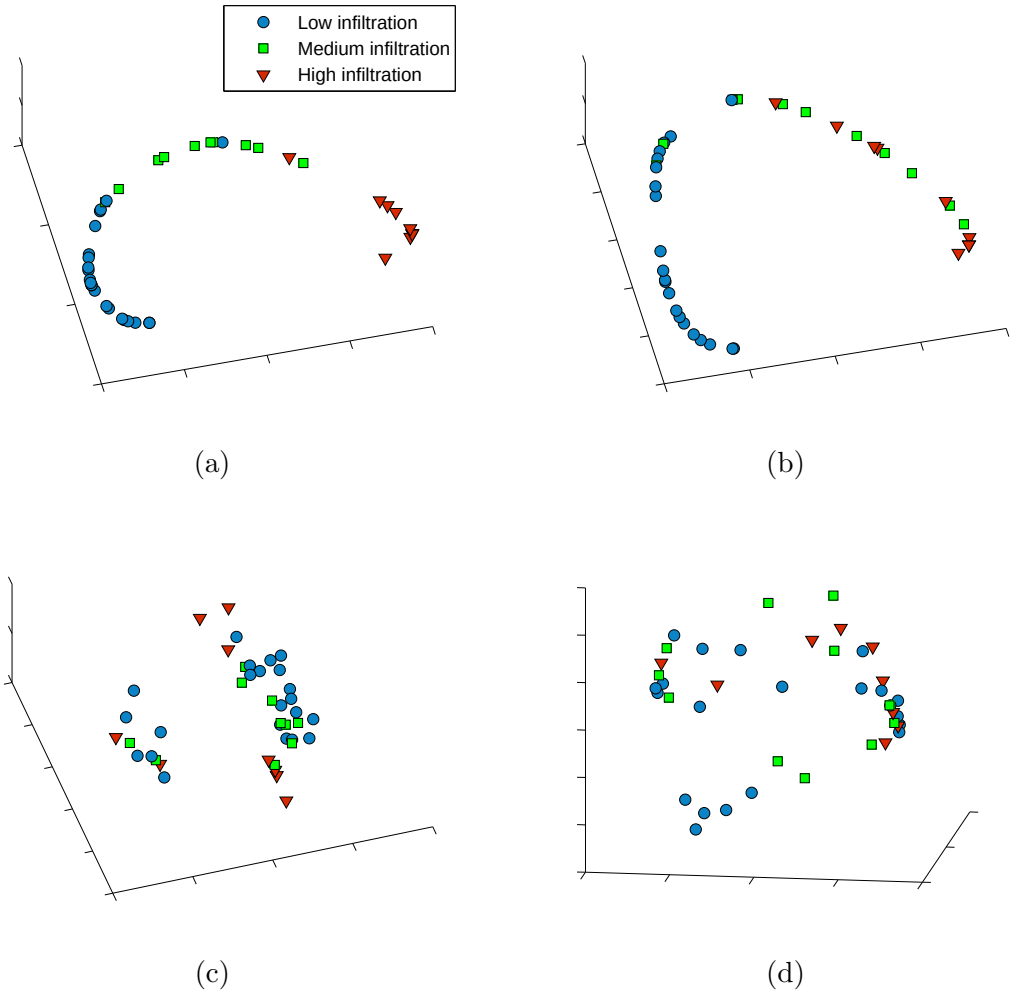

Figure 5.3: All 41 images plotted in the Graph Embedding (GE) reduced 3-dimensional Eigen space for the architectural feature set derived from (a) manual and (b) automated lymphocyte detection. Embeddings of the (c) Varma-Zisserman features with $K = 3$ and (d) Gabor filter features are also shown. The 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.

Chapter 6

Concluding Remarks and Directions for Future Research

The primary objective of this thesis is to develop an automated CADx system for detecting and stratifying the extent of LI in digitized BC histopathology. The three specific aims addressed are the ability to:

- Aim 1: Automatically detect LI in digitized BC histopathology images,
- Aim 2: Extract image-based features to quantify LI extent, and
- Aim 3: Visualize the stratification of LI extent on a low-dimensional data manifold.

The CADx system has demonstrated the ability to isolate LI from the surrounding BC nuclei, stroma, and baseline level of lymphocytes using a region-growing algorithm followed by an MRF-based refinement. Additionally, the architectural (graph-based and nuclear) features were found to be more successful than textural (VZ) features in distinguishing LI extent. Furthermore, non-linearly reducing the high-dimensional architectural image feature space reveals the presence of a smooth, continuous manifold on which BC samples are arranged with progressively increasing LI extent. While applying Graph Embedding to the high-dimensional feature space allowed for the visualization of a smooth data manifold, it did not adversely affect the classification accuracy of the SVM classifier. 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 automated LI detection algorithm.

In future research, the work presented in this thesis can be extended in two major directions. First, due to the clinical importance of LI in HER2+ BC, the ability of this CADx algorithm to stratify LI extent into low, medium, and high grades may have translational significance, whereby a prognostic test could be developed for predicting disease outcome and patient survival. Second, since the methods presented in this thesis are generalizable, the CADx system could be developed into a framework for the characterization of LI extent in other tissues and diseases.

References

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer statistics, 2009." *CA Cancer J Clin*, vol. 59, no. 4, pp. 225–249, 2009. [Online]. Available: http://dx.doi.org/10.3322/caac.20006

[2] F. Bertucci and D. Birnbaum, "Reasons for breast cancer heterogeneity." *J Biol*, vol. 7, no. 2, p. 6, 2008.

[3] S. Aaltomaa, P. Lipponen, 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.

[4] S. Demaria, M. D. Volm, R. L. Shapiro, H. T. Yee, R. Oratz, S. C. Formenti, F. Muggia, and W. F. Symmans, "Development of tumor-infiltrating lymphocytes in breast cancer after neoadjuvant paclitaxel chemotherapy." *Clin Cancer Res*, vol. 7, no. 10, pp. 3025–3030, Oct 2001.

[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] A. Rody, U. Holtrich, L. Pusztai, C. Liedtke, R. Gaetje, E. Ruckhaeberle, C. Solbach, L. Hanker, A. Ahr, D. Metzler, K. Engels, T. Karn, and M. Kaufmann, "T-cell metagene predicts a favorable prognosis in estrogen receptor-negative and her2-positive breast cancers." *Breast Cancer Res*, vol. 11, no. 2, p. R15, 2009. [Online]. Available: http://dx.doi.org/10.1186/bcr2234

[7] J. P. Monaco, J. E. Tomaszewski, M. D. Feldman, M. Moradi, P. Mousavi, A. Boag, C. Davidson, P. Abolmaesumi, and A. Madabhushi, "Detection of prostate cancer from whole-mount histology images using markov random fields," in *Workshop on Microscopic Image Analysis with Applications in Biology (in conjunction with MICCAI)*, 2008. [Online]. Available: http://www.miaab.org/miaab-2008-papers/28-miaab-2008-paper-22.pdf

[8] S. Hojjatoleslami and J. Kittler, "Region growing: a new approach," *IEEE Trans. on Image Processing*, vol. 7, no. 7, pp. 1079–1084, 1998.

[9] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, November 1984.

[10] J. Besag, "Statistical analysis of dirty pictures," *Journal of Royal Statistic Society*, vol. B, no. 68, pp. 259–302, 1986.

[11] 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 International Symposium on Biomedical Imaging: From Nano to Macro*, 2008, pp. 496–499.

[12] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "Automated grading of prostate cancer using architectural and textural image features," in *Proc. 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2007, pp. 1284–1287.

[13] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[14] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[15] G. Lee, C. Rodriguez, and A. Madabhushi, "Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 3, pp. 368–384, 2008.

[16] Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis." *Acad Radiol*, vol. 6, no. 1, pp. 22–33, Jan 1999.

[17] T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center." *Radiology*, vol. 220, no. 3, pp. 781–786, Sep 2001.

[18] 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.

[19] 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.

[20] 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 Imaging*, vol. 6, p. 14, 2006.

[21] B. Karaali and A. Tzeren, "Automated detection of regions of interest for tissue microarray experiments: an image texture analysis." *BMC Med Imaging*, vol. 7, p. 2, 2007.

[22] C. Gunduz, B. Yener, and S. H. Gultekin, "The cell graphs of cancer." *Bioinformatics*, vol. 20 Suppl 1, pp. i145–i151, Aug 2004.

[23] 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 a membrane isolation algorithm and quantitative analysis of positive controls." *BMC Med Imaging*, vol. 8, p. 11, 2008.

[24] 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.

[25] F. Schnorrenberg, C. Pattichis, K. Kyriacou, and C. Schizas, "Computer-aided detection of breast cancer nuclei," *IEEE Trans. on Information Technology in Biomedicine*, vol. 1, no. 2, pp. 128–140, 1997.

[26] P. Bamford and B. Lovell, "Unsupervised cell nucleus segmentation with active contours," *Signal Processing*, vol. 71, no. 2, pp. 203–213, 1998.

[27] 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.

[28] 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 International Symposium on Biomedical Imaging: From Nano to Macro ISBI 2008*, 2008, pp. 284–287.

[29] 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," *International Journal of Functional Informatics and Personalised Medicine*, vol. 1, pp. 26–52(27), 2008.

[30] 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.

[31] 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 *IEEE Conference on Bioinformatics and Bioengineering (BIBE)*, 2009.

[32] A. Basavanhally, S. Agner, G. Alexe, G. Bhanot, S. Ganesan, and A. Madabhushi, "Manifold learning with graph-based features for identifying extent of lymphocytic infiltration from high grade, her2+ breast cancer histology," in *Workshop on Microscopic Image Analysis with Applications in Biology (in conjunction with MICCAI)*, 2008. [Online]. Available: http://www.miaab.org/miaab-2008-papers/27-miaab-2008-paper-21.pdf

[33] 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.

[34] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, no. 6, pp. 610–621, Nov. 1973.

[35] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int J Comput Vision*, vol. 62, no. 1-2, pp. 61–81, 2005.

[36] 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.

[37] O. Tuzel, L. Yang, P. Meer, and D. J. Foran, "Classification of hematologic malignancies using texton signatures," *Pattern Analysis & Applications*, vol. 10, no. 4, pp. 277–290, 2007.

[38] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[39] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2322, 2000.

[40] M. W. Schwarz, W. B. Cowan, and J. C. Beatty, "An experimental comparison of rgb, yiq, lab, hsv, and opponent color models," *ACM Trans. on Graphics*, vol. 6, no. 2, pp. 123–158, 1987.

[41] R. Duda, P. Hart, and D. Stork, *Pattern Classification.* John Wiley & Sons, 2001.

[42] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, None, Ed. McGraw Hill, Inc., 1965.

[43] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[44] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[45] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications." *IEEE Trans Med Imaging*, vol. 21, no. 12, pp. 1552–1563, Dec 2002. [Online]. Available: http://dx.doi.org/10.1109/TMI.2002.806569

[46] J. Naik, S. Doyle, A. Basavanhally, S. Ganesan, M. Feldman, J. Tomaszewski, and A. Madabhushi, "A boosted distance metric: Application to content based image retrieval and classification of digitized histopathology," in *SPIE Medical Imaging*, vol. 7260, 2009. [Online]. Available: http://dx.doi.org/10.1117/12.813931

Curriculum Vita

Ajay Basavanhally

Education

2007-2010 M.S. in Biomedical Engineering, Rutgers University, Piscataway, NJ.

2003-2007 B.S. in Biomedical Engineering, Case Western Reserve University, Cleveland, OH.

Publications

Peer-reviewed Journal Papers

- Fatakdawala, H., Basavanhally, A., Xu, J., Ganesan, S., Feldman, M., Tomaszewski, J., Madabhushi, A., Expectation Maximization driven Geodesic Active Contour with Overlap Resolution (EMaGACOR): Application to Lymphocyte Segmentation on Breast Cancer Histopathology, IEEE Transactions on Biomedical Engineering 2010 (in press).

- Basavanhally, A., Ganesan, S., Agner, S., Monaco, J., Feldman, M., Tomaszewski, J., Bhanot, G., Madabhushi, A., Computerized Image-Based Detection and Grading of Lymphocytic Infiltration in HER2+ Breast Cancer Histopathology, IEEE Transactions on Biomedical Engineering 2010 (in press).

- Alexe, G., Monaco, J., Doyle, S., Basavanhally, A., Reddy, A., Seiler, M., Ganesan, S., Bhanot, G., Madabhushi, A., Towards Improved Cancer Diagnosis and Prognosis using Analysis of Gene Expression Data and Computer Aided Imaging, Proc Soc Exp Biol Med 2009 0: 0902-MR-89.

- Jenkins, M.W., Chughtai, O.Q., Basavanhally, A.N., Watanabe, M., and Rollins, A.M., In vivo gated 4D imaging of the embryonic heart using optical coherence tomography, J. Biomed. Opt. 12, 030505 (2007).