COMPUTERIZED DETECTION, SEGMENTATION AND CLASSIFICATION OF DIGITAL PATHOLOGY: CASE STUDY IN PROSTATE CANCER

BY SCOTT DOYLE

A dissertation 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

Degree of Doctor of Philosophy

Graduate Program in Biomedical Engineering

Written under the direction of

Anant Madabhushi

and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2011

ABSTRACT OF THE DISSERTATION

Computerized Detection, Segmentation and Classification of Digital Pathology: Case Study in Prostate Cancer

by Scott Doyle

Dissertation Director: Anant Madabhushi

Digital pathology refers to the use of scanning hardware and viewing software to digitize samples of stained pathological tissue excised from a patient. Image analysis algorithms can be employed to assist in analyzing these digital samples, increasing the speed and efficiency with which pathology samples are examined in the clinic. Traditionally these algorithms have focused on simple quantification (e.g. cell counting or stain enhancement), but the most recent developments have focused on developing quantitative disease signatures for different tissue types.

In this dissertation, an image analysis framework for automated interpretation of histology samples is described using novel image descriptors and new classification techniques. This interpretation of samples has several advantages over the traditional method of manual analysis: (1) by using quantitative disease metrics, it can be applied in a standardized fashion across several institutions with perfect agreement; (2) advanced pattern recognition and machine intelligence algorithms such as supervised classification, intelligent training, and content-based image retrieval can be employed to add to the information used to make a decision regarding diagnosis and treatment; and (3) by providing such an in-depth analysis of tissue, we can make predictions regarding the potential outcome of patients with respect to specific treatment regiments.

The overall goal of the framework is to reduce the burden on pathologists who must examine hundreds of thousands of tissue images every year, and to enhance the ability of clinicians to detect, diagnose, and treat disease.

We apply our framework to a series of datasets with a focus on detection, segmentation, and classification of prostate cancer. Our data consists of over 100 patient biopsy samples stained with hematoxylin and eosin and digitized at 40x optical magnification. Ground truth for normal, diseased, and confounder tissue was manually applied by expert pathologists. We demonstrate the ability of our framework to perform the following: detect suspicious regions of tissue on whole biopsies; analyze those regions in detail using morphological, textural, and architectural characteristics to correctly classify each region; provide an intelligent method for training the classifier and identifying new tissue classes; and perform content-based image retrieval of images in the database.

Preface

This thesis represents the collective published and unpublished works of the author. Chapters 2-9 are primarily composed of the contents of conference papers [1, 2, 3, 4, 5, 6, 7], and peer-reviewed journal articles [8], written by the author of this dissertation over the course of his thesis work.

Acknowledgements

Thanks to the members of the Laboratory for Computational Imaging and Bioinformatics at the Department of Biomedical Engineering at Rutgers University. Your feedback, collaboration, help, and guidance were essential in the completion of this work.

Thanks to my thesis advisor, Dr. Anant Madabhushi, whose patient guidance and firm insistence on excellence has led this work to where it stands today.

Finally, many thanks to the members of my thesis committee, who have dedicated precious time and effort to oversee this work, critique its weaknesses, and guide it to completion. This would not have been possible without your assistance.

Dedication

To Maxine, for sharing a life with me

To Tim and Susan, for their friendship and love

To Mark and Machiko, for their kind words and advice

To Tom and Joan, for their unconditional support and encouragement

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Digital pathology refers to the use of scanning hardware and viewing software to digitize samples of stained pathological tissue excised from a patient. In recent years, both academic and clinical centers are taking advantage of the benefits of digital pathology, including digital storage of patient cases, remote viewing of samples (telepathology), and increasingly sophisticated image analysis algorithms. Traditionally these algorithms have focused on simple quantification (e.g. cell counting or stain enhancement), but the most recent developments have focused on developing quantitative disease signatures for different tissue types.

In this dissertation, an image analysis framework for automated interpretation of histology samples is described using novel image descriptors and new classification techniques. The goal of the framework is to reduce the burden on pathologists who must examine hundreds of thousands of tissue images every year, and to enhance the ability of clinicians to detect, diagnose, and treat disease.

The system is described in the following chapters.

In Chapter 2, pathology images scanned and digitized at 20x optical magnification are decomposed into an image pyramid, providing a multi-resolution image representation. Quantitative image descriptors based on image texture (including first- and second-order statistical features and steerable filter features) are calculated in a pixel-wise fashion from the smallest resolution level, and a boosted Bayesian classifier is trained on a labeled training set of cancerous and non-cancerous pixels. Pixels classified as cancerous are subsequently analyzed at the next-largest resolution level. On a database of 100 prostate biopsy images from 58 patients, we achieve an area under the receiver operating characteristic curve (AUC) of 0.84 at the smallest resolution. At

higher resolutions, individual pixels represent too small of an area for tissue analysis, so a region-based approach was necessary.

Following cancer detection, in Chapters 3 and 4, we developed an intelligent method of training a classifier in the case where (1) a "target" class (e.g. cancerous tissue) represents a small percentage of the overall problem domain, and (2) labeled data are expensive to obtain in terms of time, money, or annotation effort. The domain of digital pathology exhibits both of these characteristics, as the "target" class of disease is typically a minority class, and annotation of samples for classifier training is very expensive. We developed a Class-Balanced Active Learning (CBAL) method of training a supervised classifier that identifies "informative" samples as those from an unlabeled dataset that are difficult to classify, and thus require manual labeling. By training on only these informative samples, we can achieve an improvement of 2-3% in accuracy and area under the receiver-operating characteristic curve (AUC) compared with classifiers trained on the same number of samples obtained with random sampling. Subsequently, we developed an advanced form of active learning termed the Consensus of Ambiguity approach, which fuses the results of multiple active learning algorithms to further reduce the size of the samples required for annotation.

In Chapter 5, we describe a Gleason grading system based on region-based features using textural, architectural, and morphological characteristics. Textural features are similar to those calculated in the first section, based on image statistics and steerable filter response. Architectural features are based on the density, arrangement, and relationship of nuclei, calculated from graph representations of the nuclei in an image. Morphological features are based on the size and shape of gland structures in the tissue. We tested these features in two domains. In the first, 54 regions of prostate biopsy tissues were classified as epithelium, stroma, Gleason grade 3, or Gleason grade 4 using a support-vector machine (SVM) operating on each pair of tissue classes and achieved an accuracy of 92.4% in separating normal tissues (epithelium vs. stroma), 76.9% in separating Gleason grade 3 from grade 4, and between 85.4% and 92.8% in separating cancerous from non-cancerous tissue types. In the second study, 48 images of breast biopsy tissue were classified as cancerous vs. non-cancerous with an accuracy of 95.8%,

and low vs. high Bloom-Richardson cancer grades with 93.3% accuracy.

In addition, in Chapter 6, we describe a content-based image retrieval framework for prostate cancer tissue. Dimensionality reduction algorithms can project data from a high- to low-dimensional space where image similarity can easily be computed between two images. By computing the similarity between an unlabeled "query" image and a database of images from a known set of classes in this low-dimensional representative space, we can retrieve similar images on the basis of quantitative features as opposed to traditional text-based labels. Using the textural, architectural, and morphological features calculated above, our system was able to compare and retrieve database images of prostate biopsy samples from Gleason grades 3 and 4 as well as epithelial tissues.

In Chapter 7, we employ a cascaded approach for true multi-class classification to correctly identify confounding tissue classes (those that mimic other classes of interest). Our multi-class approach operates using a series of binary classifiers that operate on domain-specific "class groups" for the purpose of maximizing class separability and minimizing class heterogeneity. Our cascaded approach was tested on regions of prostate biopsy tissue to distinguish normal (epithelium and stroma), cancerous (Gleason grades 3, 4, and 5), and confounder (atrophy and prostatic intraepithelial neoplasia) classes, and was compared with previous multi-class approaches including one-shot classification (OSC) and a one-versus-all (OVA) strategy. The cascaded approach was able to achieve an increase in performance (accuracy and positive predictive value) of over 24% compared with traditional multi-class approaches.

Finally, Chapters 8 and 9 describe two additional experiments performed during the course of this thesis work that make use of a number of techniques developed herein. Chapter 8 describes a set of experiments where we compared the effect of JPEG2000 compression on a gland-detection-based classification system for histopathology, and we compared the results of classification (in terms of accuracy) against the manual analysis of the tissue samples (in terms of pathologist confidence in diagnosis). These experiments are important in developing a commercially-viable system that can perform telepathology – transmission of digitized pathology across the Internet – and for long-term storage of digitized samples. Chapter 9 refers to experiments performed on

a breast cancer dataset, while maintaining several of the features that were critical in distinguishing tissue samples for prostate. We performed cancer vs. non-cancer classification as well as low vs. high degrees of malignancy of the tissue, defined by the Bloom-Richardson grading scheme.

Chapter 2

Detection of Prostate Cancer from Digitized Histopathology

2.1 Abstract

Diagnosis of prostate cancer (CaP) currently involves examining tissue samples for CaP presence and extent via a microscope, a time-consuming and subjective process. With the advent of digital pathology, computer-aided algorithms can now be applied to disease detection on digitized glass slides. The size of these digitized histology images (hundreds of millions of pixels) presents a formidable challenge for any computerized image analysis program. In this paper, we present a boosted Bayesian multi-resolution (BBMR) system to identify regions of CaP on digital biopsy slides. Such a system would serve as an important preceding step to a Gleason grading algorithm where the objective would be to score the invasiveness and severity of the disease. In the first step, our algorithm decomposes the whole-slide image into an image pyramid comprised of multiple resolution levels. Regions identified as cancer via a Bayesian classifier at lower resolution levels are subsequently examined in greater detail at higher resolution levels, thereby allowing for rapid and efficient analysis of large images. At each resolution level, 10 image features are chosen from a pool of over 900 first-order statistical, second-order co-occurrence, and Gabor filter features using an AdaBoost ensemble method. The BBMR scheme, operating on 100 images obtained from 58 patients, yielded (1) areas under the receiver operating characteristic curve (AUC) of 0.84, 0.83, and 0.76 at the lowest, intermediate, and highest resolution levels respectively, and (2) an 8-fold savings in terms of computational time compared to running the algorithm directly at full (highest) resolution. The BBMR model out-performed (in terms of AUC) (1) individual features (no ensemble) and (2) a random forest classifier ensemble obtained

by bagging multiple decision tree classifiers. The apparent drop-off in AUC at higher image resolutions is due to lack of fine detail in the expert annotation of CaP and is not an artifact of the classifier. The implicit feature selection done via the AdaBoost component of the BBMR classifier reveals that different classes and types of image features become more relevant for discriminating between CaP and benign areas at different image resolutions.

2.2 Introduction

American Cancer Society predicts that over 192,000 new cases of prostate cancer (CaP) will be diagnosed in the United States in 2009, and over 27,000 men will die due to the disease. Successful treatment for CaP depends largely on early diagnosis, determined via manual analysis of biopsy samples [12]. Over one million prostate biopsies are performed annually in the US, each of which generates approximately 6-14 tissue samples. These samples are subsequently analyzed for presence and grade of disease under a microscope by a pathologist. Approximately 60-70% of these biopsies are negative for CaP [13], implying that the majority of a pathologist's time is spent examining benign tissue. Regions identified as CaP are assigned a Gleason score, reflecting the degree of malignancy of the tumor based on the patterns present in the sample [11]. Accurate tissue grading is impeded by a number of factors, including pathologist fatigue, variability in application and interpretation of grading criteria, and the presence of benign tissue that mimics the appearance of CaP (benign hyperplasia, high-grade prostatic intraepithelial neoplasia) [14, 15]. These pitfalls can be mitigated by introducing a quantitative "second reader" capable of automatically, accurately, and reproducibly finding suspicious CaP regions on the image [16]. Such a system would allow the pathologist to spend more time determining the grade of the cancerous regions and less time on finding them.

The recent emergence of "digital pathology" has necessitated work on developing quantitative and automated computerized image analysis algorithms to assist pathologists in interpreting the large quantities of digitized histological image data being generated via whole slide digital scanners [17]. Computer aided diagnosis (CAD) algorithms

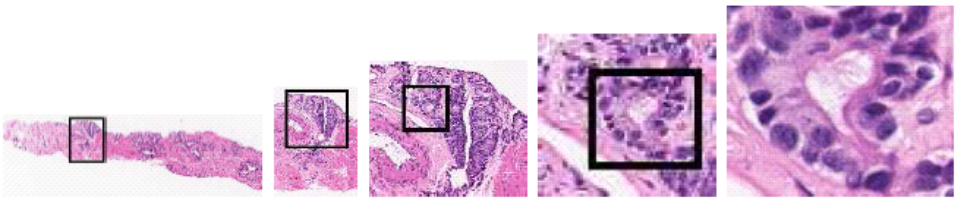


Figure 2.1: Illustration of the multi-resolution approach, where lower resolutions are used to identify suspicious regions that are later analyzed at higher resolution. This multi-resolution approach results in significant computational savings. The most discriminatory features for CaP detection are learned and used to train a classifier at each image resolution.

have been proposed for detecting neuroblastoma [18], identifying and quantifying extent of lymphocytic infiltration on breast biopsy tissue [19], and grading astrocytomas in brain biopsies [20]. In the context of detecting prostate cancer on histopathology, previous CAD approaches have employed low-level image characteristics such as color, texture, and wavelets [21], second-order statistical [22], and morphometric attributes [23] in conjunction with classifier systems to distinguish benign from CaP regions. Diamond, et al. [24] devised a system for distinguishing between stroma, benign epithelium, and prostate cancer images measuring 100×100 pixels in size taken from whole-mount histology specimens. Using morphological and texture features, an overall accuracy of 79.3% was obtained on 8,789 samples, each of which represented a homogeneous section of tissue. Tabesh, et al. [23] presented a CAD system for distinguishing between (1) 367 CaP and non-CaP regions, and (2) 268 images of low and high Gleason grades of CaP on tissue microarray images using texture, color, and morphometric features, achieving an accuracy of 96.7% and 81.0% for each respective task. However, these results only reflect the system accuracy when distinguishing between small spots on a tissue microarray. Farjam, et al. [25] used size and shape of gland structures in selected regions of prostate tissue to determine the corresponding Gleason grade of the cancer. An average accuracy of 96.5% in correctly classifying the Gleason grade (1-5) of two different sets of images was obtained. Again, these results are acheived on pre-selected image regions, where the implicit assumption was that the tissue was homogeneous across the region of interest (ROI).

One of the most challenging tasks in developing CAD algorithms for grading disease

on digitized histology is to first easily identify the spatial extent and presence of disease which can then be subjected to a more detailed analysis [25, 26, 27]. The reliance on pre-selected ROIs limits the general usability of the automated grading algorithms, since ROI determination is not a trivial problem; one may argue even more challenging than grading pre-extracted ROIs. Ideally, a comprehensive CAD algorithm would first detect these suspicious ROIs in a whole-slide image; the image having been digitized at high optical magnification (generating images with millions of pixels that take up several gigabytes of hardware memory). Once these ROIs have been identified, a separate suite of grading algorithms can be leveraged to score the invasiveness and malignancy of the disease in the ROIs. In this paper, we address the former problem of automatically detecting CaP regions from whole slide digital images of biopsy tissue quickly and efficiently, allowing the pathologist to focus on a more detailed analysis of the cancerous region for the purposes of grading.

Our methodology employs a boosted Bayesian multi-resolution (BBMR) classifier to identify suspicious areas, in a manner similar to an expert pathologist who will typically examine the tissue sample via a microscope at multiple magnifications to find regions of CaP. Figure 2.1 illustrates the scheme employed in this work for CaP detection by hierarchically analyzing the image at multiple resolutions. The original image obtained from a scanner is decomposed into successively lower representations to generate an "image pyramid." Low resolutions (near the "peak" of the pyramid) are analyzed rapidly. A classifier trained on image features at the lowest resolution is used to assign a probability of CaP presence at the pixel level. Based on a pre-defined threshold value, obviously benign regions are eliminated at the lowest resolution. Successive image resolutions are analyzed in this hierarchical fashion until a spatial map of disease extent is obtained, which can then be employed for Gleason grading. This approach is inspired by the use of multi-resolution image features employed by Viola and Jones [28], where coarse image features were used to rapidly identify ROIs for face detection, followed by computationally expensive but detailed features calculated on those ROIs. This led to an overall reduction in the computational time for the algorithm. For our study, we begin with low-resolution images that are fast to analyze but contain little

structural detail. Once obviously benign areas are eliminated, high-resolution image analysis of suspicious ROIs is performed.

At each resolution level, we perform color normalization by converting the image from the red, green, and blue (RGB) color space to the hue, saturation, and intensity (HSI) space to mitigate variability in illumination caused by differences in scanning, staining, or lighting of the biopsy sample. From each of these channels, we extract a set of image features extracted at the pixel level that include first-order statistical, second-order co-occurrence [29], and wavelet features [30, 31]. The rationale for these texture features is twofold. (1) First- and second-order texture statistics mitigate the sensitivity of the classifier to variations in illumination and color. (2) It is known that cancerous glands in the prostate tend to be arranged in an arbitrary fashion, so that in CaP dominated regions, the averaged gland orientation is approximately zero. In normal areas, glands tend to be strongly oriented in a particular direction. The choice of wavelet features (e.g. Gabor) is dictated by the desire to exploit the differences in orientation of structures in normal and CaP regions. At low resolution levels, it is expected that subtle differences in color and texture patterns between the CaP and benign classes, captured by first- and second-order image statistics, will be important for class discrimination, whereas at higher resolution levels when the orientation and size of individual glands become discernible, wavelet and orientation based features [31] will be more important (Figure 2.1).

Kong, et al. [18] employed a similar multi-resolution framework for grading neuroblastoma on digitized histopathology. They were able to distinguish three degrees of differentiation in neuroblastoma with an overall accuracy of 87.88%. In that study subsets of features obtained via sequential floating forward selection were subjected to dimensionality reduction and tissue regions were classified hierarchically using a weighted combination of nearest-neighbor, nearest-mean, Bayesian, and support vector machine (SVM) classifiers. During this process, the meaning of the individual features is lost through the dimensionality reduction and classifier combination. Sboner, et al. [32] used a multi-classifier system to determine whether an image of a skin lesion corresponds to melanoma or a benign nevus using either an "all-or-none" rule, where all

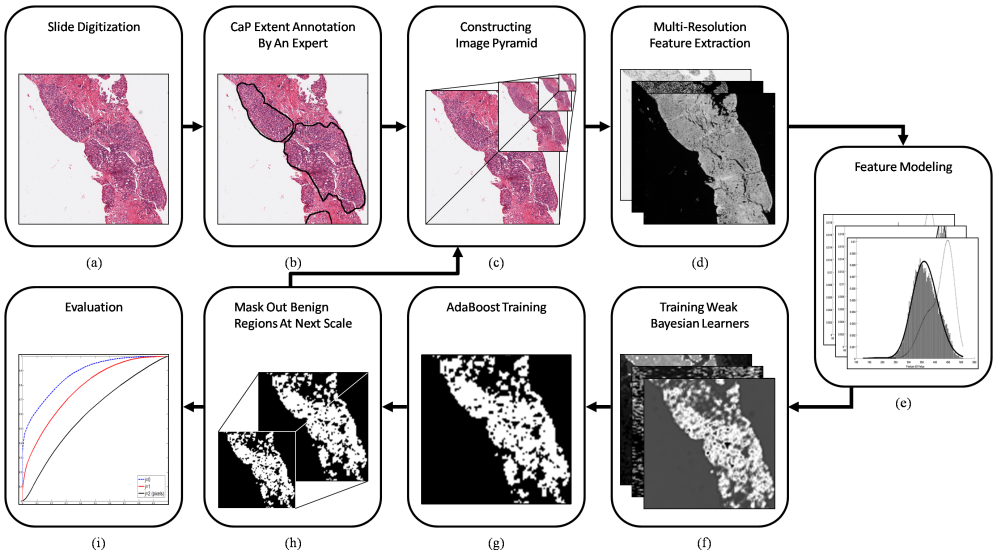


Figure 2.2: Flowchart illustration of the working of the BBMR algorithm. (a) Slide digitization captures tissue samples at high resolution, and (b) ground truth regions of cancer are manually labeled. (c) Pyramidal decomposition is performed to obtain a set of successively smaller resolution levels. (d) At each level, several image features are extracted and (e) modeled via a Bayesian framework. (f) The weak classifiers thus constructed are combined using (g) the AdaBoost algorithm [9] into a single strong classifier for a specific resolution. (h) The probabilistic output of the AdaBoost classifier [9] is then converted to a hard output reflecting the extent of the CaP region (based on the operating point of the ROC curve learned during training). Thus obviously benign regions are masked out at the next-highest resolution level. The process repeats until image resolution is sufficient for application of advanced region-based grading algorithms. (i) Evaluation is performed against the expert-labeled ground-truth.

classifiers must agree that a lesion is benign for it to be classified as such, or a "majority voting" rule, where two out of three classifiers is taken as the final result. However, this set of rules is based on a number of domain-specific assumptions and is not suitable for high-dimensional feature ensembles. Hinrichs, et al. [33] employed Linear Programming boosting (LPboosting) where a linear optimization approach is taken to combine multiple features; however, the LP approach does not provide a clear insight on feature ranking or selection, and it is difficult to derive an intuitive understanding of why certain features out-perform others. Madabhushi, et al. [34] evaluated 14 different classifier ensemble schemes for the purpose of detecting prostate cancer in images of high resolution *ex vivo* MRI, showing that the technique used to create ensembles and the relevant parameters can have an effect on the resulting classificaiton performance, given identical training and testing data.

In our work, we have sought to select and extract features in a way that reflects

Data Set	Sample Size	Image Sizes	Parameters	Ground Truth	Performance Measures
H&E stained prostate tissue	58 patient studies (100 images)	Roughly 50,000 pixels/dimension (original)	40X optical magnification	Manual annotation	Pixel accuracy, ROC curve

Table 2.1: Description of the dataset, image parameters, ground truth annotation, and performance measures used in this study.

visual image differences in the cancer and benign classes at each image resolution. To that end, we model the extracted features in a Bayesian framework to generate a set of weak classifiers, which are combined using a set of feature weights determined via the AdaBoost algorithm [9]. Each feature's weight is determined by how well the feature can discriminate between cancer and non-cancer regions, enabling implicit feature selection at each resolution by choosing the features with the highest weights. The computational expense involved in training the AdaBoost algorithm is mitigated by the use of the multi-resolution scheme. In our scheme, the classifier allows for connecting the performance of a feature to physical or visual cues used by pathologists to identify malignant tissue, thereby providing an intuitive understanding as to why some features can discriminate between tissue types more effectively than others. A similar task was performed by Ochs, et al. [35], who employed a similar AdaBoost technique to the classification of lung bronchovascular anatomy in CT. In that study, AdaBoost-generated feature weights provided insight into how different features performed in terms of their discriminative capabilities, an important characteristic in designing and understanding a biological image classification system. Unlike ensemble methods that sample the feature space (random forests) [36] or project the data into higher dimensional space (SVMs) [37], the AdaBoost algorithm provides a quantitative measurement of which features are important for accurate classification, thus providing a look at which features are providing the discriminatory information used to distinguish the cancer and non-cancer classes.

Our methodology, called the boosted Bayesian multi-resolution (BBMR) approach, has two main advantages: (1) it can identify suspicious tissue regions from a whole-slide scan of a prostate biopsy as a precursor to automated Gleason grading; and (2)

it can process large images quickly and quantitatively, providing a framework for rapid and standardized analysis of full biopsy samples at high resolution. We quantitatively determine the efficiency of our methodology with respect to different classifier ensembles on a set of 100 biopsy images (image sizes range from 10-50 thousand pixels along each dimension) taken from 58 patient studies.

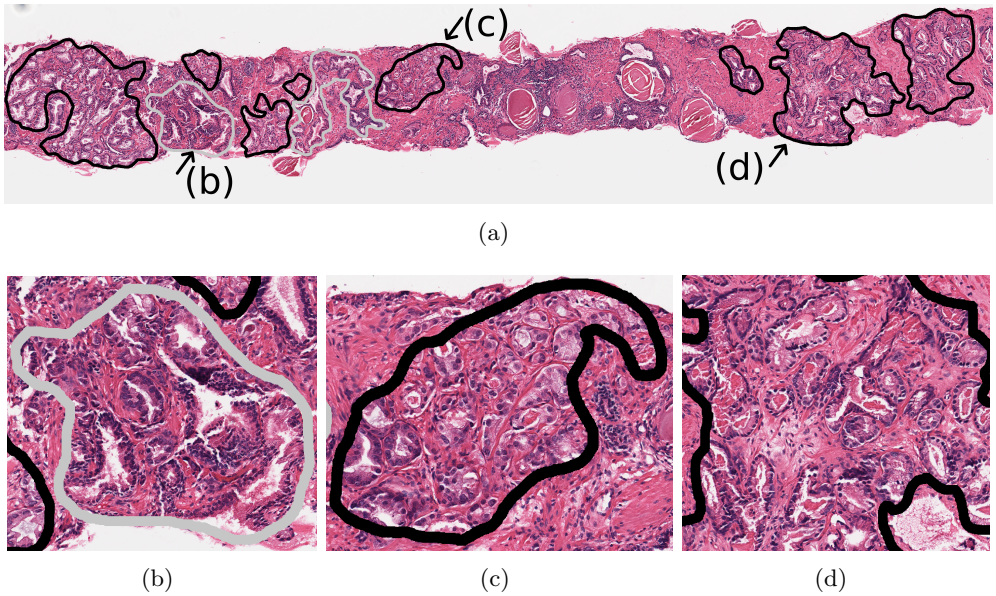
The rest of this paper is organized as follows. In Section 2 we discuss our data set and the initial preprocessing steps. In Section 3 we discuss the feature extraction procedure. In Section 4 we describe the BBMR algorithm. Experimental design is described in Section 5, and the results of analysis are presented in Section 6. Discussion of the results and concluding remarks are presented in Sections 7 and 8, respectively.

2.3 Brief Overview of Methodology and Preprocessing of Data

2.3.1 Image Digitization and Decomposition

An overview of our methodology is illustrated in Figure 2.2. A cohort of 100 human prostate tissue biopsy cores taken from 58 patients are fixed onto glass slides and stained with hematoxylin (H) and eosin (E) to visualize cell nuclei and extra- and intracellular proteins. The glass slides are then scanned into a computer using a ScanScope CS$^{\mathrm{TM}}$whole-slide scanning system operating at 40× optical magnification. Images are saved to disk using the ImageScope$^{\mathrm{TM}}$software package as 8-bit tagged image file format (TIFF) files (scanner and software both from Aperio, Vista, CA). Tissue staining, fixing, and scanning were done at the Department of Surgical Pathology at the University of Pennsylvania. The images digitized at the 40× magnification ranged in size from 10,000 to 50,000 pixels along each of the X and Y axes, depending on the orientation and size of the tissue sample on a slide, with file sizes ranging between 1-2 gigabytes.

An image pyramid was created using the pyramidal decomposition algorithm described by Burt and Adelson [38]. In this procedure, Gaussian smoothing is performed on the full resolution (40×) image followed by sub sampling of the smoothed image by a factor of two. This reduces the image size to one-half of the original height and width; the process is repeated n times to generate an image pyramid of successively smaller



(a)

(b) (c) (d)

Figure 2.3: Shown are: (a) the original image with cancer (black contours) and non-cancer (gray contour) regions labeled by an expert, (b) closeup of the non-cancer region, and (c), (d) closeups of cancerous regions. Regions shown in (b), (c), and (d) are indicated on (a) by black arrows and text labels.

and lower resolution images. The value of n depends on the structures in the image; a large n corresponds to several different image resolutions. A summary of the data is given in Table 2.1.

2.3.2 Color Normalization

Variations in illumination caused by improper staining or changes in ambient lighting conditions at the time of digitization may dramatically affect image characteristics, potentially affecting classifier performance. To deal with this potential artifact, we convert the images from the original RGB (red, green, blue) color space captured by the scanner to the HSI (hue, saturation, intensity) space. In the HSI space, intensity or brightness in a channel are kept separate from the color information. This will confine variation in brightness and illumination to only one channel (intensity), whereas the RGB space combines brightness and color [39]. Thus differences that naturally occur between different biopsy slides will be constrained to one channel instead of affecting all three.

Symbol	Description	Symbol	Description
$\mathcal{C} = (C, f)$	Image scene	$\mathcal{P}$	n-level Image pyramid
$\mathbf{F}(c)$	Feature vector	Φ_u	Random variable for u
$P(\omega_i \vert f_u(c))$	Probability of class ω_i	$p(f_u(c) \vert \omega_i)$	PDF of class ω_i
O_c	c co-occurrence matrix	$N_w(c)$	Windowsize w around c
κ, θ	Gabor filter parameters	$\mathbf{G}$	Gabor filter function
Γ	Gamma function	τ, η	Gamma parameters
$\Pi^{\text{Ada}}(c)$	Classification; $\{0, 1\}$	$\Pi_u(c)$	Weak classifier for c
T	Num. of weak classifiers	δ_u	Classifier threshold
$\alpha_1, \cdots, \alpha_T$	Weak classifier weights	$h_1, \cdots, h_T$	Weak classifiers
$\mathcal{G} = (C, g)$	Ground truth scene	$\mathcal{D}$	AdaBoost distribution
$\mathcal{A}(c)$	Ensemble result for c	$\mathcal{B}^j$	Classification scene

Table 2.2: List of frequently appearing notation and symbols in this paper.

2.3.3 Ground Truth Annotation for Disease Extent

For each of the 100 images used in this study, ground truth labels were manually assigned by an expert pathologist using the ImageScope$^{\text{TM}}$slide-viewing software. Labels were placed on the original scanned image and were propagated through the pyramid using the decomposition procedure described in Section 2.3.1. The expert was instructed to label all cancer within the tissue image for training and evaluation purposes and was permitted to use any magnification necessary to accurately delineate CaP spatial extent. A subset of the non-cancer class, comprising benign epithelium and stroma, was also labeled for training; for evaluation, all non-cancer regions (whether labeled as benign or unlabeled) were considered to be benign. Regions where both cancer and non-cancerous tissues appear growing in a mixed pattern were labeled as cancerous with the understanding that some stroma or benign epithelium may be contained within the cancer-labeled region (Figure 2.3(d)).

2.3.4 Notation

The notation used in this paper is summarized in Table 2.2. We represent a digitized image by a pair $\mathcal{C} = (C, f)$, where C is a 2D grid of image pixels and f is a function that assigns a value to each pixel $c \in C$. The pyramidal representation of the original image $\mathcal{C}$ is given by $\mathcal{P} = \{\mathcal{C}^0, \mathcal{C}^1, \cdots, \mathcal{C}^{n-1}\}$, where $\mathcal{C}^j = (C^j, f)$ corresponds to the image at the j-th level of the pyramid, where $j \in \{0, 1, \cdots, n-1\}$. We define the lowest (i.e. "coarsest") resolution level as $\mathcal{C}^0$, and the highest resolution level (at which

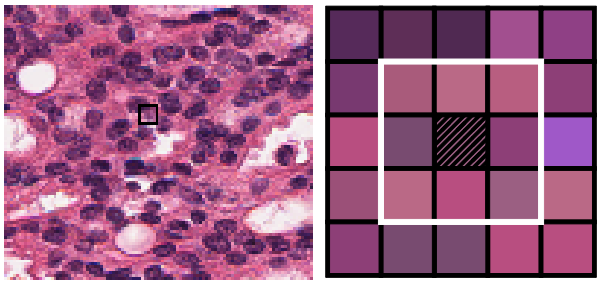
Feature Class and Individual Attributes	Parameters	Total Features
First-order Statistics	Window size: $w \in \{3, 5, 7\}$	135
Co-occurrence Features	Window size: $w \in \{3, 5, 7\}$ Distance: $\Delta = 1$	144
Gabor Features	Window size: $w \in \{3, 5, 7\}$ Frequency: $\kappa \in \{0, 1, \cdots, 7\}$ Phase: $\theta \in \{\frac{\pi}{8}, \frac{2\pi}{8}, \cdots, \frac{8\pi}{8}\}$	648

Table 2.3: Summary of the features used in this study, including a breakdown of each of the three major feature classes (first-order, second-order Haralick, and Gabor filter) with associated filter parameters. Also included (right column) are the total number of features calculated for each feature class.

the image was originally scanned) as $\mathcal{C}^{n-1}$. For brevity, notation referring to pyramidal level is only included when such a distinction is necessary. At each resolution level, feature extraction is performed such that for each pixel $c \in C$ in an image, we obtain a K-dimensional feature vector $\mathbf{F}(c) = [f_u(c)|u \in \{1, 2, \cdots, K\}]$, where $f_u(c)$ is the value of feature u at pixel $c \in C$. We denote as Φ_u, where $u \in \{1, 2, \cdots, K\}$, the random variable associated with each of the K features. An observation of Φ_u is made by calculating $f_u(c)$, for $c \in C$.

2.4 Feature Extraction

The operations described below are performed on a neighborhood of pixels, denoted N_w, centered on the pixel of interest where w denotes the radius of the neighborhood. This is illustrated in Figure 2.4. At every $c \in C$, $N_w(c) = \{d \in C | d \neq c, ||d - c||_\infty \leq w\}$, where $|| \cdot ||_\infty$ is the L_∞ norm. Feature value $f_u(c)$ is calculated on the values of the pixels in $N_w(c)$. This is done for all pixels in an image which yields the corresponding feature image. For a single pixel, $c \in C$, the K-dimensional feature vector is denoted by $\mathbf{F}(c)$. Some representative feature images are shown in Figure 2.5. The black contour in Figure 2.5(a) represents the cancer region. Table 2.3 summarizes the image features extracted; details regarding the computation of the individual feature classes are given below.



(a) (b)

Figure 2.4: Illustration of the procedure for calculating image features. (a) Magnified region of the original tissue image, (b) pixel-wise magnification of the region with the window N_w ($w = 3$) indicated by a white border and center pixel shaded with diagonal stripes.

First-order Statistics

A total of 135 first-order statistical features are calculated from each image. These features included average, median, standard deviation, and range of the image intensities within small neighborhoods centered at every image pixel. Additionally, Sobel filters in the X, Y, and 2 diagonal axes, 3 Kirsch filter features, gradients in the X and Y axes, difference of gradients, and diagonal derivative for window sizes $w \in \{3, 5, 7\}$ were also extracted.

Co-occurrence Features

Co-occurrence features [29] are computed by constructing a symmetric 256×256 co-occurrence matrix, O_c, for each $N_w(c)$, $c \in C$, where O_c describes the frequency with which two different pixel intensities appear together within a fixed neighborhood. The number of rows and columns in the matrix O_c are determined by the maximum possible intensity value in the image, I. For 8-bit images, I corresponds to $2^8 = 256$. The value $O_c[a, b]$ for $a, b \in \{1, \cdots, I\}$ represents the number of times two distinct pixels, $d, k \in N_w(c)$, with pixel values $f(d) = a$ and $f(k) = b$, are within a unit distance of each other. A detailed description of the construction of O_c can be found in [29]. From O_c a set of Haralick features (Joint Entropy, Energy, Inertia, Inverse Difference Moment, Correlation, two Measurements of Correlation, Sum Average, Sum Variance, Sum Entropy, Difference Average, Difference Variance, Difference Entropy, Shade, Prominence,

and Variance) are extracted. These 16 features are calculated from each of the three image channels (hue, saturation, and intensity) for $w \in \{3, 5, 7\}$, yielding a total of 144 co-occurrence image features.

Steerable Filters

The Gabor filter is constructed as a Gaussian function modulated by a sinusoid [31, 40]. The filter provides a large response for image regions with intensity patterns that match the filter's orientation and frequency shift parameters. For a pixel $c \in C$ located at image coordinates (x, y), the Gabor filter bank response is given as:

$$\mathbf{G}(x, y, \theta, \kappa) = e^{-\frac{1}{2}((\frac{x'}{\sigma_x})^2+(\frac{y'}{\sigma_y})^2)} \cos(2\pi\kappa x'), \tag{2.1}$$

where $x' = x\cos(\theta)+y\sin(\theta)$, $y' = y\cos(\theta)+x\sin(\theta)$, κ is the filter's frequency shift, θ is the filter phase, σ_x and σ_y are the standard deviations along the X, Y axes. We created a filter bank using eight different frequency-shift values $\kappa \in \{0, 1, \cdots, 7\}$ and nine orientation parameter values ($\theta = \frac{\epsilon\cdot\pi}{8}$ where $\epsilon \in \{0, 1, \cdots, 8\}$), generating 72 different filters. The response for each of these was calculated for window sizes $w \in \{3, 5, 7\}$ and from each of the three image channels (hue, saturation, and intensity), yielding a total of 648 Gabor features.

2.5 Boosted Bayesian Multi-Resolution (BBMR) Classifier

2.5.1 Bayesian Modeling of Feature Values

For each image feature extracted (see Section 2.4), a training set of labeled samples is employed to construct a *probability density function* (PDF) $p(f_u(c)|\omega_i)$, which is the likelihood of observing feature value $f_u(c)$ for class ω_i, where $u \in \{1, 2, \cdots, K\}$, $i \in \{1, 0\}$. We refer to the cancer class as ω_1 and the non-cancer class as ω_0. The posterior class-conditional probability that pixel c belongs to class ω_i is denoted as $P(\omega_i|f_u(c))$ and may be obtained via Bayes Rule [41]. In this study a total of $K = 927$ PDFs are generated, one for each of the extracted texture features.

The PDFs are modeled in the following way. For each random variable Φ_u, for $u \in \{1, 2, \cdots, K\}$, we are interested in modeling the *a posteriori probability*, denoted

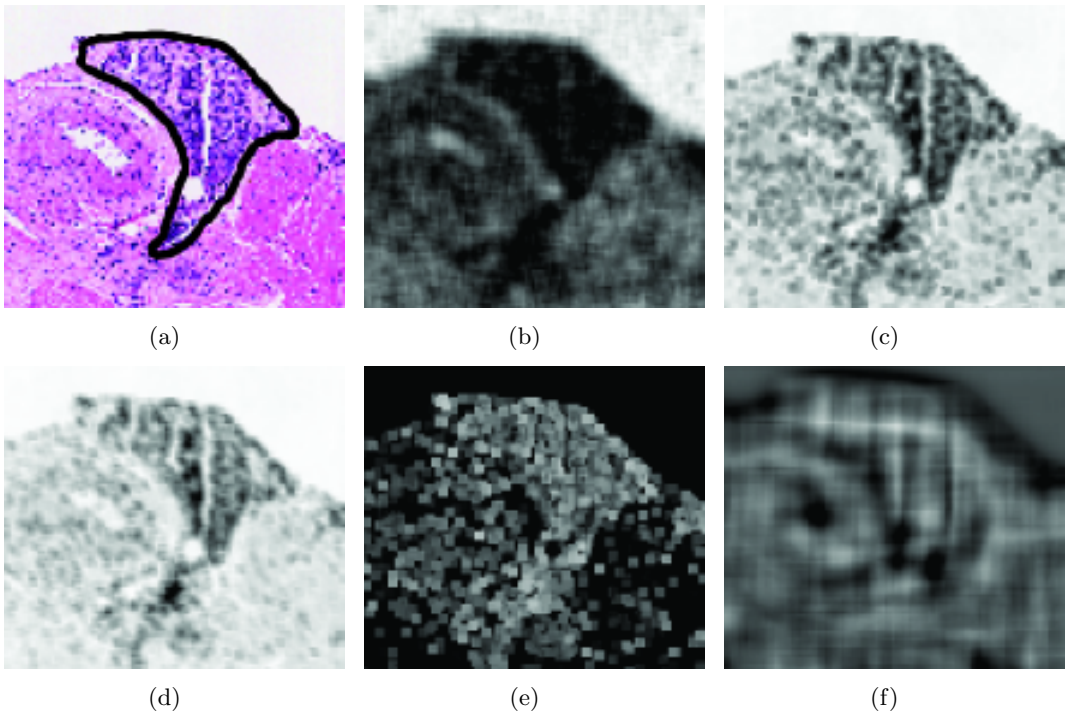


Figure 2.5: (a) Original digitized prostate histopathological image with the manual segmentation of cancer overlaid (black contour), and 5 corresponding feature scenes: (b) correlation ($w = 7$), (c) sum variance ($w = 3$), (d) Gabor filter ($\theta = \frac{5 \cdot \pi}{8}$, $\kappa = 2$, $w = 3$), (e) difference ($w = 3$), and (f) standard deviation ($w = 7$).

by $P(\omega_i|\Phi_u)$, that feature values in Φ_u reflect class ω_i. This probability is given by the Bayes Rule [41]:

$$P(\omega_i|\Phi_u) = \frac{P(\omega_i)p(\Phi_u|\omega_i)}{\sum_{k=0}^{1} P(\omega_k)p(\Phi_u|\omega_k)}, \tag{2.2}$$

where $P(\omega_k)$ is the prior probability of class ω_k and $p(\Phi_u|\omega_i)$ is the *class-conditional probability density* for ω_i given Φ_u. We can estimate the PDF as a gamma function parameterized by a scale parameter τ and a shape parameter η from the training data:

$$p(\Phi_u|\omega_i) \approx \Phi_u^{\tau-1} \frac{e^{-\Phi_u/\eta}}{\eta^\tau \Gamma(\tau)}, \tag{2.3}$$

where Γ is the gamma function and parameters $\tau, \eta > 0$. The gamma distribution was chosen over alternatives such as the Gaussian distribution due to the observed shapes of feature histograms, which tend to be asymmetric about the mean. Illustrated in Figure 2.6 are examples of parameterized PDFs corresponding to class ω_1 at resolution levels (a) $j = 0$, (b) $j = 1$, and (c) $j = 2$, as well as ω_0 at levels (d) $j = 0$, (e) $j = 1$,

and (f) $j = 2$ for the Haralick variance feature. The solid black line indicates the gamma distribution estimate, calculated from the feature values plotted as the gray histogram. Note that while the gamma distribution (Equation 2.3) models the cancer class distribution very accurately, some discrepancy between the model fit and the data for the non-cancer class is observable in Figures 2.6(d)-(f). This discrepancy between the model and the empirical data is due to the high degree of variability and heterogeneity found in the non-cancer class. Because all tissue data not labeled as cancer is considered part of the non-cancer class, the non-cancer class includes a diverse array of tissue types including stroma, normal epithelium, low- and high-grade prostatic intraepithelial neoplasia, atrophy, and inflammation [16, 42]. These diverse tissue types cause a high degree of variability in the non-cancer class, decreasing the goodness of the fit to the model. In an ideal scenario, each of these tissue types would constitute a separate class with its own model; unfortunately this is a non-trivial task, limited by the time and expense required to obtain detailed annotations of these tissue classes.

2.5.2 Boosting Weak Classifiers

We first construct a set of weak Bayesian classifiers, one for each of the extracted features, using Equation 2.2. Note that the term "weak classifier" is used here to denote a classifier constructed using a single attribute. The pixel-wise Bayesian classifier Π_u, for $c \in C$, $u \in \{1, 2, \cdots, K\}$, is constructed as:

$$\Pi_u(c) = \begin{cases} 1 & \text{if } P(\omega_1 | f_u(c)) > \delta_u, \\ 0 & \text{if } P(\omega_1 | f_u(c)) < \delta_u, \end{cases} \tag{2.4}$$

where $\Pi_u(c) = 1$ corresponds to a positive (cancer) classification, $\Pi_u(c) = 0$ corresponds to a negative (non-cancer) classification, and $\delta_u \in [0, 1]$ is a feature-specific threshold value. The optimal threshold value was learned off line on a set of training images using Otsu's thresholding method [43], a rapid method for choosing the optimal threshold by minimizing intra-class variance.

Once the weak classifiers, Π_u, for $u \in \{1, 2, \cdots, K\}$, have been constructed they

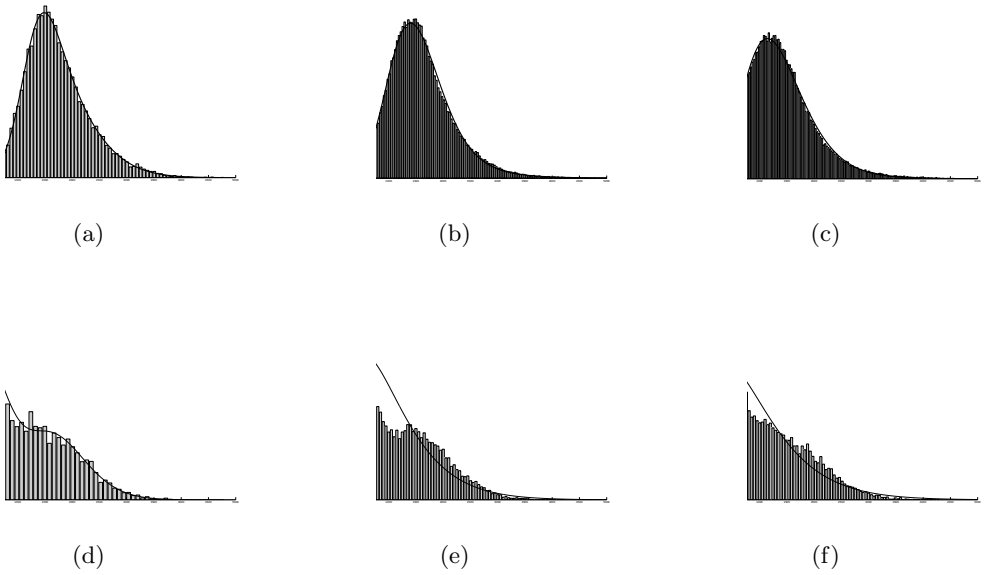


Figure 2.6: Probability density functions for the Haralick variance feature for $w = 7$. Shown are PDFs for resolutions levels (a), (d) $j=0$, (b), (e) $j=1$, and (c), (f) $j=2$. All PDFs in the top row ((a), (b), (c)) are calculated for the cancer class, and in the bottom row ((d), (e), (f)) for the non-cancer class. The best fit gamma distribution models are superimposed (black line) on the empirical data (shown in gray). The change in PDFs across different image resolution levels ($j \in \{0, 1, 2\}$) reflects the different class discriminatory information present at different resolution levels in the image pyramid.

are combined to create a single strong classifier via the AdaBoost ensemble method [9]. The output of selected classifiers is combined as a weighted average to generate the final strong classifier output. The algorithm maintains a set of weights, $\mathcal{D}$, for each of the training samples which is iteratively updated to choose classifiers that correctly classify "difficult" samples (i.e. samples that are often misclassified). The algorithm is run for T iterations to output (1) a modified set of T pixel-wise classifiers $h_1, h_2, \cdots, h_T$, where $h_1(c) \in \{1, 0\}$ indicates the output of the highest-weighted classifier, and (2) T associated weights $\alpha_1, \alpha_2, \cdots, \alpha_T$ for each classifier. Note that $\alpha_1, \alpha_2, \cdots, \alpha_T$ reflect the importance of each of the individual features (classifiers) in discriminating CaP and non-CaP areas across different image resolutions. While T is a free parameter ($1 \leq T \leq K$) it is typically chosen such that the difference in accuracy using $T + 1$ classifiers is negligible. For this study, we set $T = 10$. The result of the ensemble classifier at a given pixel c and at a specific image resolution is denoted as:

$$\mathcal{A}^{\text{Ada}}(c) = \sum_{t=1}^{T} \alpha_t h_t(c). \tag{2.5}$$

The output of the ensemble result can be thresholded to obtain a combined classification for pixel $c \in C$:

$$\Pi^{\text{Ada}}(c) = \begin{cases} 1, & \text{if } \mathcal{A}^{\text{Ada}}(c) > \delta_{\text{Ada}}, \\ 0, & \text{otherwise}, \end{cases} \tag{2.6}$$

where δ_{Ada} is chosen using Otsu's method. For additional details on the AdaBoost algorithm, we refer the reader to [9].

2.5.3 Multi-Resolution Implementation

The multi-resolution framework is illustrated in Algorithm $BBMR()$. Once the classification results are obtained from the final ensemble and at a particular image resolution, we obtain a binary image $\mathcal{B}^j = (C^j, \Pi^{\text{Ada}})$, representing the hard segmentation of CaP at the pixel level. Linear interpolation is then applied to $\mathcal{B}^j$ to resize the classification result to fit the size of the image at pyramid level $j+1$. We begin the overall multi-resolution algorithm with $j = 0$. While we construct image pyramids with $n = 7$, the classifier is only applied at image levels 0, 1, and 2. At lower image resolutions, benign and suspicious areas become difficult to resolve and at resolutions $j > 2$, significant incremental benefit is not obtained from a detection perspective. At higher image resolutions, the clinical problem is more about the grading of the invasiveness of the disease and not about detection. Note that in this work we are not addressing the grading problem.

2.6 Experiments and Evaluation Methods

2.6.1 Experimental Design

Our system was evaluated on a total of 100 digitized tissue sample images obtained from 58 patients. We evaluated the classification performance of the BBMR system using (a) qualitative likelihood scene analysis, (b) area under the receiver operating

Algorithm BBMR()
Input: Image pyramid $\mathcal{P} = \{\mathcal{C}^0, \mathcal{C}^1, \cdots, \mathcal{C}^{n-1}\}$
Output: Binary output at final resolution level $\mathcal{B}^{n-1}$
begin
0. initialize $\mathcal{B}^0$ to include all pixels in C^0
1. *for* $j = 0$ to 2 *do*
2. Extract $\mathbf{F}(c)$ for non-zero pixels $c \in C^j$ and in $\mathcal{B}^j$;
3. Estimate $P(\omega_i|\Phi_u)$ for all $u \in \{1, 2, \cdots, K\}$;
4. Construct Π_u for all u;
5. Obtain Π^{Ada} via AdaBoost;
6. Obtain binary mask $\mathcal{B}^j$;
7. Resize $\mathcal{B}^j$ to obtain $\mathcal{B}^{j+1}$
8. *endfor*
end

characteristic (ROC) curve, and (c) classification accuracy at the pixel level (Section 2.6.3). Additional experiments were performed to explore different aspects of the BBMR algorithm. The list of experiments is as follows.

Experiment 1: Evaluation of BBMR Classifier

We evaluated the output of the BBMR algorithm using the metrics listed in Section 2.6.3, which include both qualitative and quantitative performance measures.

Experiment 2: Classifier Comparison

We compared the BBMR classifier, denoted as Π^{BBMR}, with five other classifiers (summarized in Table 2.4). Two aspects of the system were altered to obtain the additional classifiers. (1) The method of constructing the feature PDFs was changed from a Gamma distribution estimate (Section 2.5) to a non-parametric PDF obtained directly from the feature histograms (the gray bars in Figure 2.6). (2) The method used for combining weak classifiers was changed from the AdaBoost method to a randomized forest ensemble method [36]. Additionally, we tested a non-ensemble approach where the single best-performing feature was used for classification. Different combinations

PDF Construction	Ensemble Method		
	AdaBoost	Random Forests	Best Feature
Parametric	Π^{BBMR}	$\Pi^{\text{RF, gamma}}$	$\Pi^{\text{best,gamma}}$
Non-Parametric	$\Pi^{\text{Ada,feat}}$	$\Pi^{\text{RF,feat}}$	$\Pi^{\text{best,feat}}$

Table 2.4: List of the different classifiers compared in this work. The BBMR algorithm is denoted as Π^{BBMR}, while the additional classifiers are denoted with their feature estimation method (parametric (gamma) distribution or non-parametric feature distribution), as well as the different ensemble methods (AdaBoost, Random Forests, or None (single feature).

of the method for generating the PDFs (parametric and non-parametric) and ensembles (AdaBoost, Random Forests) yield the five additional classifiers shown in Table 2.4. While many additional ensemble and classification methods exist (for example, extremely randomized trees [44] is a recent alternative to random forests, and the support vector machine [37] is a classifier that does not employ a probability density function), it is beyond the scope of this paper to empirically test all combinations of these methodologies. The purpose of testing the five classifiers in Table 4 is to show that BBMR can provide similar or better performance compared to some other common ensemble based classifier schemes, in addition to the other benefits of transparency and speed.

Experiment 3: BBMR Parameter Analysis

We evaluated three aspects of the BBMR scheme: (1) the number of weak classifiers used in the ensemble, T; (2) types of features selected at each image resolution level; and (3) the computational savings of using the BBMR approach.

2.6.2 Classifier Training

BBMR Classifier

To ensure robustness of BBMR to training data, randomized 3-fold cross-validation was performed on both a per-image and a per-patient basis.

Image-Based Cross-Validation: Cross-validation is performed on a per-image basis since images taken from the same patient are assumed to be independent. This is motivated by the fact that biopsy cores are obtained in a randomized fashion from

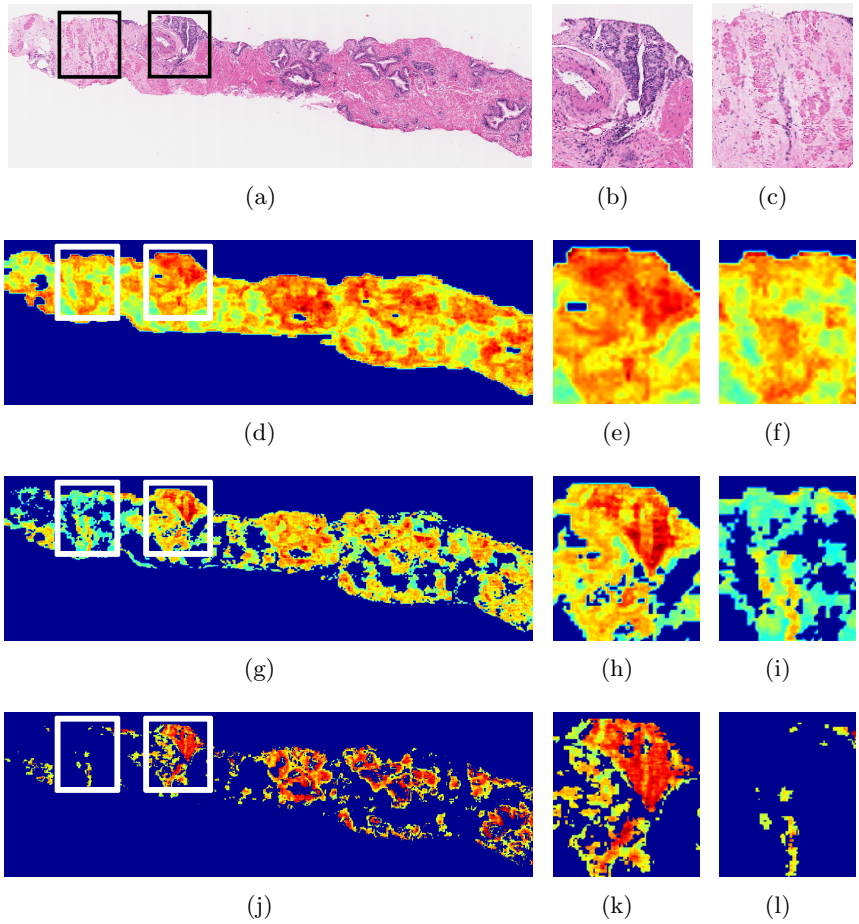


Figure 2.7: An illustration of CaP classification via Π^{BBMR} on a single prostate image sample. The full image is shown in (a), with corresponding likelihood scenes $\mathcal{L}^0$, $\mathcal{L}^1$, and $\mathcal{L}^2$ shown in (d), (g), and (j), respectively. Closeups of cancer and benign regions (indicated by boxes on the full image) are shown in (b) and (c), respectively, with corresponding CaP classification shown in subsequent rows as for the full image. Note the decrease in false positive classifications (third column) compared to the stability of the cancerous regions in the second column.

different sextant locations within the prostate, and the appearance of cancer regions within a single patient can be highly heterogeneous. Thus, for the purposes of finding pixels that contain cancer, each image is independent. The entire dataset (100 images) is randomly split into three roughly equal groups: G_1, G_2, and G_3, each representing a collection of images. Each trial consists of three rounds: In the first round, the classifier is trained using pixels drawn at random from images in groups G_1 and G_2, and is tested by classifying the pixels from images in G_3. The purpose of sampling pixels at random is to ensure that equal numbers of cancer and non-cancer pixels are selected for training. For testing, all pixels in the image are used for testing, except for

those left out as a result of non-cancer classification at lower resolution levels. In the second round, G_1 and G_3 are used to generate the training, and the pixels from images in G_2 are classified. Here, the PDFs are re-created using features calculated from pixels in the G_1 and G_3 groups. As before, equal numbers of cancer and non-cancer samples are used to generate the training set, while all of the pixels in G_2 that have not been classified as non-cancer at a previous scale are used for testing. In the third and final round, G_2 and G_3 are used for training, and G_1 is classified. In this way, each image has its pixels classified exactly once per trial using training pixels drawn at random from two-thirds of the images in the dataset. The second trial repeats the process, randomly generating new G_1, G_2, and G_3 sets of images. A total of 50 trials were run in this manner to ensure classifier robustness to training data.

Patient-Based Cross-Validation: In addition, we performed a second set of cross-validation experiments where G_1, G_2, and G_3 contain images from separate patients; that is, a single patient could not have images in more than one group, ensuring that training images were from different patients than testing images.

Random Forest Classifier

The random forest ensemble constructs a set of decision trees using a random subset of training data and a randomly chosen set of features. The output of each tree represents a binary classification "vote" for that tree, and the number of trees in the ensemble determines the maximum number of votes. Hence, each random forest classifier yields a fuzzy voting scene. The voting scene is thresholded for some value δ_{RF} (determined via Otsu's method), yielding a binary scene. The random forest ensemble is evaluated using accuracy and area under the ROC curve as described in Section 2.6.3. A total of T trees were used in the ensemble, each of which used a maximum of $\frac{K}{T}$ randomly-selected features to generate the tree. For each of the trees, the C4.5 algorithm [45] was employed to construct and prune the tree. Each tree was optimally pruned to a length that helped maximize classifier accuracy.

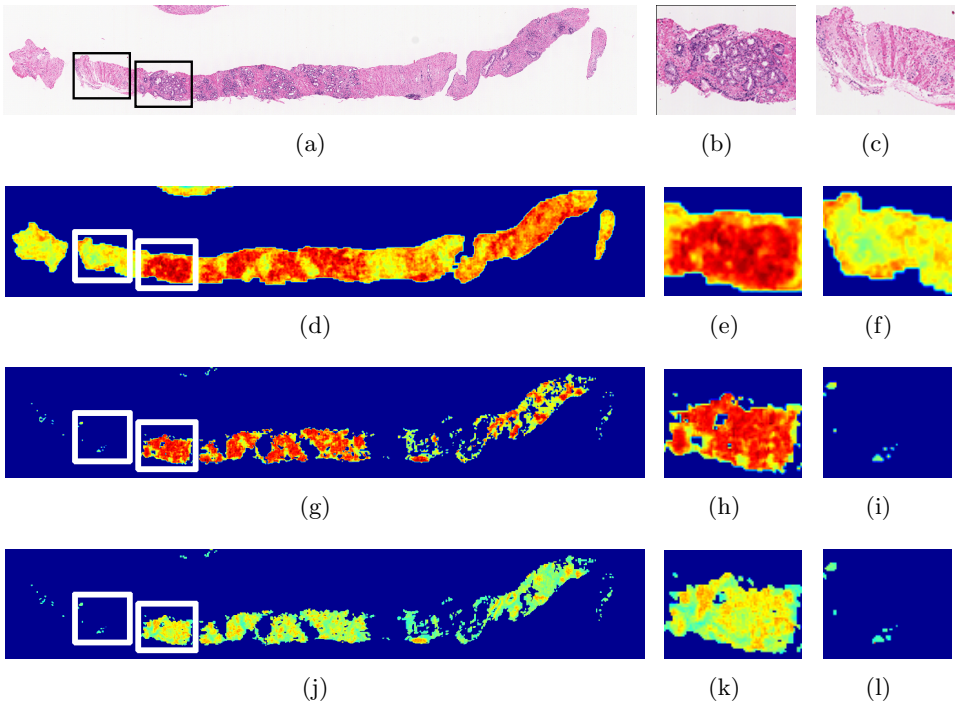


Figure 2.8: An illustration of CaP classification via Π^{BBMR} on a single prostate image sample. The full image is shown in (a), with corresponding likelihood scenes $\mathcal{L}^0$, $\mathcal{L}^1$, and $\mathcal{L}^2$ shown in (d), (g), and (j), respectively. Closeups of cancer and benign regions are shown in (b) and (c), with corresponding CaP classification shown in subsequent rows as for the full image.

2.6.3 Evaluation Methods

Evaluation of classification performance is done at the pixel level. The number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) pixels was determined for each image. We denote the expert manually-labeled ground truth for tumor in $\mathcal{C}$ as $\mathcal{G} = (C, g)$, where $g(c) = 1$ for cancer and $g(c) = 0$ for non-cancer pixels, for all $c \in C$. We determine three methods for classifier evaluation: (a) likelihood scene analysis, (b) area under the ROC curve (AUC), and (c) accuracy.

Comparative Analysis of Classifier-Generated CaP Probability

We obtain a likelihood scene $\mathcal{L}^j = (C^j, \mathcal{A}^{\text{Ada}})$ for image resolution level $j \in \{0, 1, 2\}$ where each pixel's value given by $\mathcal{A}^{\text{Ada}}(c)$ (Eq. 2.5). Likelihood scenes are compared with pathologist-defined ground truth for a qualitative evaluation of classifier performance.

Area Under the ROC Curve (AUC)

Classifier sensitivity (SENS) and specificity (SPEC) in detecting disease extent is determined by varying δ_{Ada} (Section 2.5.2). For a specific threshold δ_{Ada}, and for all $c \in C$, the number of true positives ($\text{TP}_{\delta_{\text{Ada}}}$) is found as $|\{c \in C | g(c) = \Pi^{\text{Ada}}(c) = 1\}|$, false positives ($\text{FP}_{\delta_{\text{Ada}}}$) is $|\{c \in C | g(c) = 0, \Pi^{\text{Ada}}(c) = 1\}|$, true negatives ($\text{TN}_{\delta_{\text{Ada}}}$) is $|\{c \in C | g(c) = \Pi^{\text{Ada}}(c) = 0\}|$, and false negatives ($\text{FN}_{\delta_{\text{Ada}}}$) is $|\{c \in C | g(c) = 1, \Pi^{\text{Ada}}(c) = 0\}|$, where $|\mathcal{S}|$ denotes the cardinality of set $\mathcal{S}$. For brevity, we ignore notation referring to the threshold for TP, TN, FP, and FN. $\text{SENS}_{\delta_{\text{Ada}}}$ and $\text{SPEC}_{\delta_{\text{Ada}}}$ can then be determined as,

$$\text{SENS}_{\delta_{\text{Ada}}} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2.7}$$

$$\text{SPEC}_{\delta_{\text{Ada}}} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{2.8}$$

By varying the threshold as $0 \leq \delta_{\text{Ada}} \leq \max[\mathcal{A}^{\text{Ada}}]$, ROC curves for all the classifiers considered can be plotted by varying sensitivity vs. $1 -$ specificity for the full range of threshold values. A large area under the ROC curve (AUC ≈ 1.0) reflects superior classifier discrimination between the cancer and non-cancer classes.

Accuracy

The accuracy of the system at threshold δ_{Ada} is determined as:

$$\text{ACC}_{\delta_{\text{Ada}}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{\text{TP} + \text{TN}}{|C|}. \tag{2.9}$$

For our evaluation, we choose δ_{Ada} as described in Section 2.5.2, using Otsu's thresholding. The motivation for this thresholding technique as opposed to the use of the operating point of the ROC curve is that the operating point finds a trade-off between sensitivity and specificity, while we wish to favor false positives over false negatives (since a false negative would be propagated through higher resolution levels when the masks are resized).

2.7 Experimental Results

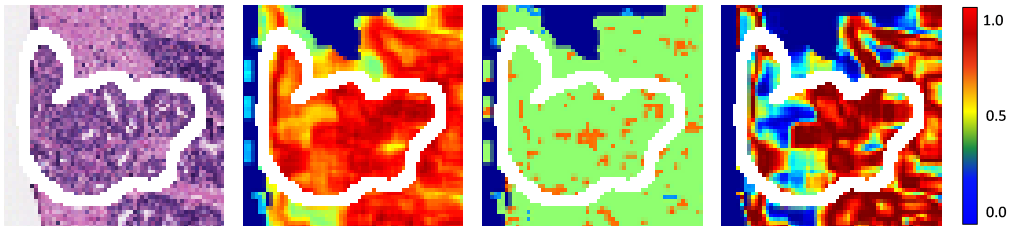
2.7.1 Experiment 1: Evaluation of BBMR Classifier

Figures 2.7 and 2.8 show qualitative results of Π^{BBMR} on two sample images from the database. The original image is shown in the top row, with the corresponding likelihood scenes $\mathcal{L}^0$, $\mathcal{L}^1$, and $\mathcal{L}^2$ shown in the second, third, and fourth rows, respectively. It is important to note that the increase in resolution levels changes the likelihood values from $j = 0$ (second row) to $j = 2$ (fourth row). Shown in Figures 2.7(b) and 2.8(b) are magnified image areas corresponding to the cancer region (as determined by the pathologist), and shown in Figures 2.7(c) and 2.8(c) are non-cancer image areas. As the image resolution increases, more benign regions are pruned and eliminated at the lower resolutions.

2.7.2 Experiment 2: Classifier Comparison

The comparison of average classifier accuracy (μ_{ACC}) and AUC (μ_{AUC}) values for each of the classifiers listed in Table 2.4 is shown for the image-based cross-validation experiments in Table 2.5. Shown in Table 2.6 are sample results for a patient-based cross-validation experiment, where images from the same patient are grouped together for cross-validation. μ_{ACC} is calculated at the threshold determined by Otsu's method, and μ_{AUC} was obtained across 50 trials with 3-fold cross-validation. Figure 2.10(a) illustrates ROC curves for the BBMR classifier over all images in our database at image resolution levels $j = 0$ (blue dashed line), $j = 1$ (red solid line), and $j = 2$ (black solid line).

In Figure 2.9 is a qualitative comparison of the three different feature ensemble methods: Figure 2.9(a) shows the original image with the cancer region denoted in white, while the BBMR, random forest, and single-feature classifiers are shown in Figures 2.9(b), (c), and (d), respectively. The displayed images are from image resolution level $j = 1$. When compared with the BBMR method, the random forest ensemble is unable to find CaP regions with high probability, while the single best feature cannot capture the entire cancer region. False negatives at this resolution level would be



(a) (b) (c) (d)

Figure 2.9: Qualitative comparison of classifier performance on a single prostate image sample. The original image is shown in (a) with the cancer region denoted in a white contour. Likelihood scenes corresponding to Π^{BBMR}, $\Pi^{\text{RF,gamma}}$, and $\Pi^{\text{best,gamma}}$ are shown in (b), (c), and (d), respectively. All images are shown from resolution level $j = 1$. The BBMR method is able to detect CaP with a higher probability than the random forest ensemble, and with fewer false negatives than when using a single feature.

propagated at the next level, decreasing overall accuracy.

	Level 0		**Level 1**		**Level 2**	
	μ_{AUC}	μ_{ACC}	μ_{AUC}	μ_{ACC}	μ_{AUC}	μ_{ACC}
Π^{BBMR}	**0.84 (0.10)**	**0.69 (0.04)**	**0.83 (0.09)**	**0.70 (0.05)**	**0.76 (0.09)**	0.68 (0.06)
$\Pi^{\text{RF,gamma}}$	0.33 (0.03)	0.46 (0.01)	0.37 (0.11)	0.50 (0.01)	0.56 (0.04)	0.31 (0.04)
$\Pi^{\text{best,gamma}}$	0.54 (0.06)	0.56 (0.10)	0.54 (0.05)	0.55 (0.11)	0.55 (0.04)	0.55 (0.14)
$\Pi^{\text{Ada,feat}}$	0.32 (0.04)	0.53 (0.05)	0.25 (0.11)	0.53 (0.05)	0.20 (0.04)	**0.69 (0.16)**
$\Pi^{\text{RF,feat}}$	0.73 (0.01)	0.50 (0.02)	0.66 (0.01)	0.52 (0.03)	0.63 (0.01)	0.60 (0.01)
$\Pi^{\text{best,feat}}$	0.44 (0.03)	0.52 (0.01)	0.26 (0.01)	0.46 (0.03)	0.43 (0.01)	0.54 (0.01)

Table 2.5: Image-based cross-validation results. Shown are ACC and AUC values for all the classifiers considered in this study (listed in Table 2.4) at each of the 3 image resolution levels. Average accuracy and AUC over all images in the database are shown with standard deviation in parentheses. The largest value in each column is shown in bold.

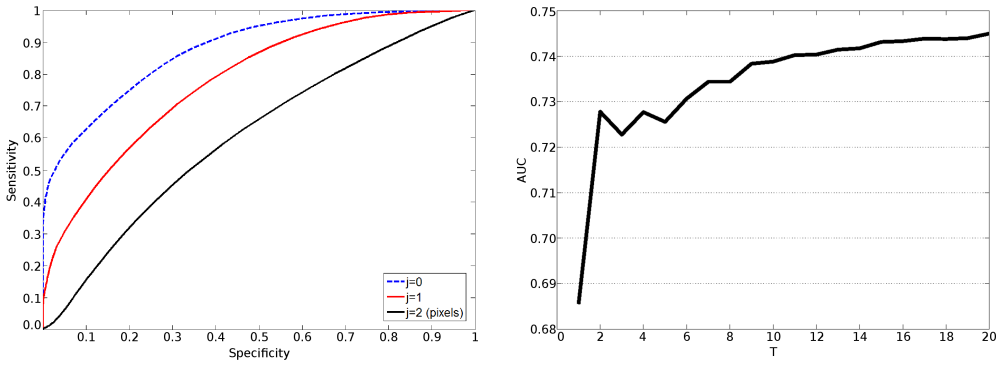
	Level 0		**Level 1**		**Level 2**	
	μ_{AUC}	μ_{ACC}	μ_{AUC}	μ_{ACC}	μ_{AUC}	μ_{ACC}
Π^{BBMR}	**0.85 (0.03)**	**0.74 (0.02)**	**0.83 (0.03)**	**0.66 (0.02)**	**0.60 (0.01)**	**0.57 (0.01)**
$\Pi^{\text{RF,feat}}$	0.79 (0.03)	0.63 (0.07)	0.73 (0.01)	0.55 (0.01)	0.59 (0.01)	0.51 (0.03)

Table 2.6: Patient-based cross-validation results. Shown are ACC and AUC values for Π^{BBMR} and $\Pi^{\text{RF,feat}}$ at each of the 3 image resolution levels. Average accuracy and AUC over all images in the database are shown with standard deviation in parentheses. The largest value in each column is shown in bold.

2.7.3 Experiment 3: BBMR Parameter Analysis

AdaBoost Ensemble Size T

The graph in Figure 2.10(b) illustrates how the AUC values for Π^{BBMR} change with T, i.e. the number of weak classifiers combined to generate a strong classifier ensemble.



(a) (b)

Figure 2.10: (a) ROC curves generated at $j = 0$ (dashed blue line), $j = 1$ (solid red line), and $j = 2$ (solid black line) using the BBMR classifier. The apparent decrease in classifier (BBMR) accuracy with increasing image resolution is due to a lack of granularity in image annotation at the higher resolutions (see Figure 2.12). (b) Change in AUC as a result of varying T in the BBMR AdaBoost ensemble for level $j = 2$. Similar trends were observed for $j = 0$ and $j = 1$.

The independent axis in Figure 2.10(b) shows the number of weak classifiers used in the ensemble, while the dependent axis shows the corresponding AUC for the strong BBMR classifier averaged over 100 studies at image resolution level $j = 2$. We note that as the number of classifiers increases, μ_{AUC} increases up to a point, beyond which adding additional weak classifiers does not significantly increase performance. For the plot shown in Figure 2.10(b), T was varied from 1 to 20, and μ_{AUC} remained relatively stable beyond $T = 10$. The trends shown in Figure 2.10(b) for $j = 2$ were also observed for $j = 0$ and $j = 1$.

AdaBoost Feature Selection

The features chosen by AdaBoost at each resolution level are specific to the information available at that image resolution. Table 2.7 shows the top five features selected by AdaBoost at each image resolution. Table 2.7 reveals that features corresponding to a larger scale (greater window size) performed better compared to smaller scale attributes (smaller window size), while first-order statistical greylevel features do not discriminate between cancer and benign areas at any resolution. The poor performance of first-order statistical features suggests that simple pixel-level intensity statistics (area, standard deviation, mode, etc.) are insufficient to explain image differences between

Resolution Level $j = 0$			Resolution Level $j = 1$			Resolution Level $j = 2$		
Rank	Feature	w	Rank	Feature	w	Rank	Feature	w
1	Haralick	7	1	Haralick	7	1	Haralick	7
2	Haralick	7	2	Gabor Filter	7	2	Gabor Filter	7
3	Haralick	5	3	Gabor Filter	7	3	Gabor Filter	3
4	Gabor Filter	3	4	Greylevel	7	4	Gabor Filter	7
5	Gabor Filter	7	5	Haralick	3	5	Haralick	5

Table 2.7: List of the top 5 features chosen by AdaBoost at the three resolution levels. The most important discriminatory attributes across all image resolutions are clearly second-order Haralick features, suggesting that the specific co-occurrence of image intensities is the most crucial signature to distinguish CaP and non-CaP areas.

cancer and benign regions. Additionally, Gabor features performed well across all image resolutions, suggesting that differences in texture orientation and phase are significant discriminatory attributes.

Computational Efficiency

Figure 2.11 illustrates the computational savings in employing the multi-resolution BBMR scheme. The non-multi-resolution based approach employs approximately 4 times as many pixel-level calculations as the BBMR scheme at image resolution levels $j = 1$ and $j = 2$. At the highest resolution level considered in this work and for images of approximately 1000×1000 pixels, the analysis of a single image requires less than three minutes on average. All computations in this study were done on an dual core Xeon 5140 2.33GHz computer with 32GB of RAM running the Ubuntu Linux operating system and MATLAB$^{\text{TM}}$software package version 7.7.0 (R2008b) (The MathWorks, Natick, Massachusetts).

2.8 Discussion

Examining the ROC curves in Figure 2.10(a), we can see that the classification accuracy appears to go down at higher image resolutions. We can explain the apparent drop-off in accuracy by illustrating BBMR classification on an example image at resolutions $j = 0$ and $j = 2$ (Figure 2.12). The annotated cancer region appears in a gray contour. Also shown are subsequent binary classification results obtained via the BBMR classifier at image resolution levels $j = 0$ and $j = 2$ (Figures 2.12(b) and (c) respectively).

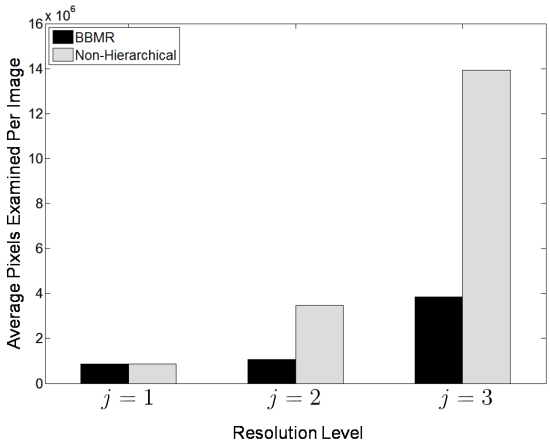
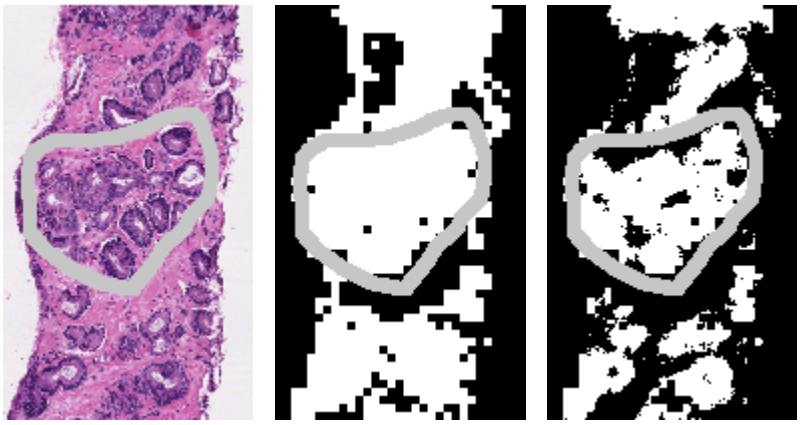


Figure 2.11: Efficiency of the system using the BBMR system (black) and a non-hierarchical method (gray), measured in terms of the number of pixel-level calculations for levels $j = 0$, $j = 1$, and $j = 2$.

As is discernible in Figure 2.12(a), the cancer region annotated by the expert is heterogeneous, comprising many benign structures (stroma, intra-gland lumen regions) within the general region. However, the manual ground truth annotation of CaP does not have the granularity to resolve these regions properly. Thus, at higher image resolutions where the pixel-wise classifier can begin to discriminate between these spurious benign regions and cancerous glands, the apparent holes within the expert delineated ground truth are reported as false negative errors. We emphasize that these reported errors are not the result of poor classifier performance; they instead illustrate an important problem in obtaining spatial CaP annotations on histology at high resolutions. At high image resolutions, a region-based classification algorithm which takes these heterogeneous structures into account is more appropriate.

The BBMR classifier was modified to perform patch-based instead of pixel-based classification at $j=2$. A uniform grid was superimposed on the original image (Figure 2.13(b)), dividing the image into uniform 30-by-30 regions. A patch was labeled as suspicious if the majority of the pixels in that patch were labeled as disease on ground truth. The BBMR classifier was trained using patches labeled as benign and diseased, since at $j \geq 2$ identifying diseased regions is more appropriate compared to pixel-based detection. The features calculated at the pixel level were averaged to provide a single value for each patch. The results of patch-wise classification on a sample image at



(a) (b) (c)

Figure 2.12: Qualitative results illustrating the apparent dropping off in pixel-level classifier accuracy at higher image resolutions. Shown in (a) is the original image with the cancer region annotated in gray. The binary image in (b) shows the overlay of the BBMR classifier detection results for CaP at $j = 0$ and (c) shows corresponding results for $j = 2$. Note that at $j = 2$, the BBMR classifier begins to discriminate between benign stromal regions and cancerous glands; that level of annotation granularity is not however captured by the expert.

level $j = 2$ are shown in Figure 2.13(d) and compared with the BBMR pixel level classifier on the same image in Figure 2.13(c). Intertwined regions of benign tissue within the diseased areas, classified as benign by the pixel-wise BBMR classifier (and labeled as "false negatives" as a result) are classified as cancerous by the patch-wise BBMR classifier. This yields an ROC curve with a greater area; compare corresponding curves for the pixel-based and patch-based BBMR classifiers in Figure 2.13(a).

We would like to point out that the apparent dropoff in classifier accuracy has to do with the lack of ground truth granularity at higher resolutions. Pathologists are able to distinguish cancer from non-cancer at low magnifications, only using higher magnifications to confirm a diagnosis and perform Gleason grading. We believe that a proper region-based algorithm, with appropriately-chosen features (such as nuclear density, tissue architecture, gland-based values, etc) will be the best method for describing tissue regions as opposed to tissue pixels, which was the objective of this work. In a Gleason grading system [23, 3], such additional high-level features will be calculated from the suspicious regions detected at the end of the BBMR classification algorithm.

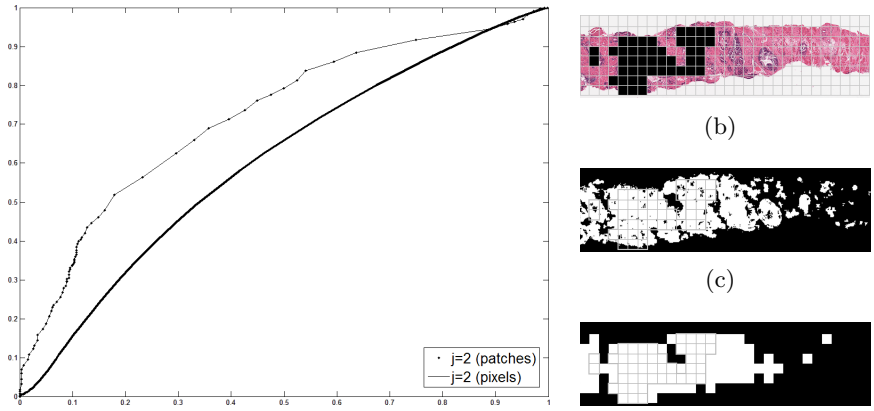


Figure 2.13: (a) Comparison of ROC curves between pixel-based classification (solid black line) and patch-based classification (black dotted line). (b) Original image with a uniform 30-by-30 grid superimposed. Black boxes indicate the cancer region. (c) Pixel-wise classification results at resolution level $j=2$, yielding the solid black ROC curve in (a). (d) Patch-wise classification results according to the regions defined by the grid in (b), yielding the dotted black ROC curve in (a). The use of patches removes spurious benign areas within the CaP ground truth region from being reported as false negatives.

2.9 Concluding Remarks

In this work, we presented a boosted Bayesian multi-resolution (BBMR) classifier for automated detection of prostate cancer (CaP) from digitized histopathology, a necessary precursor to automated Gleason grading. To the best of our knowledge this work represents the first attempt to automatically find regions involved by CaP on digital images of prostate biopsy needle cores. The classifier is able to automatically and efficiently detect areas involved by disease across multiple image resolutions (similar to the approach employed manually by pathologists) as opposed to selecting an arbitrary image resolution for analysis. The hierarchical multi-resolution BBMR classifier yields areas under the ROC curves of 0.84, 0.83, and 0.76 for the lowest, medium, and highest image resolutions, respectively. The use of a multi-resolution framework reduces the amount of time needed to analyze large images by approximately 4-6 times compared to a non-multi-resolution based approach. The implicit feature selection method via AdaBoost reveals which features are most salient at reach resolution level, allowing the classifier to be tailored to incorporate class discriminatory information as it becomes available at each image resolution. Larger scale features tended to be more informative

compared to smaller scale features across all resolution levels, with the Gabor filters (which pick up directional gradient differences) and Haralick features (which capture second order texture statistics) being the most important. We also found that the BBMR approach yielded higher AUC and accuracy than other classifiers using a random forest feature ensemble strategy, as well as those using a non-parametric formulation for feature modeling.

Pixel-wise classification breaks down as the structures in the image are better-resolved, leading to a number of "false-negative" results which are in fact correctly identified "benign" areas within the region manually delineated as CaP. This is due to a limit on the granularity of manual annotation and is not an artifact of the classifier. At high resolution, a patch-based system is more appropriate compared to pixel-level detection. The results of this patch-based classifier would serve as the input to a Gleason grade classifier at higher image resolutions.

Chapter 3

An Active Learning Based Classification Strategy for the Minority Class Problem: Application to Histopathology Annotation

3.1 Abstract

Supervised classifiers for digital pathology can improve the ability of physicians to detect and diagnose diseases such as cancer. Generating training data for classifiers is problematic, since only domain experts (e.g. pathologists) can correctly label ground truth data. Additionally, digital pathology datasets suffer from the "minority class problem", an issue where the number of exemplars from the non-target class outnumber target class exemplars. This imbalance can bias the classifier and reduce the accuracy of the predicted class labels. In this paper, we develop a training strategy that combines active learning (AL) with class-balancing. AL identifies unlabeled samples that are "informative" (i.e. likely to increase classifier performance) for annotation, avoiding non-informative samples. This yields high classifier accuracy with a smaller training set size compared with random learning (RL). Previous AL methods have not explicitly accounted for the minority class problem in biomedical images. By pre-specifying a target class ratio our classifier mitigates the problem of training bias. Finally, we develop a mathematical model to predict the number of annotations (and thus cost) required to achieve balanced classes in the training set. In addition to predicting training cost, the model reveals the theoretical properties of AL in the context of the minority class problem. Using this class-balanced AL training strategy (CBAL), we build a classifier to distinguish cancer from non-cancer regions on digitized prostate histopathology. Our dataset consists of 12,000 image regions sampled from 100 biopsies (58 prostate cancer patients). We compare CBAL against: (1) unbalanced AL (UBAL), which uses AL but

ignores class ratio; (2) class-balanced RL (CBRL), which uses RL with a specific class ratio; and (3) unbalanced RL (UBRL). The CBAL-trained classifier yields 2% greater accuracy and 3% higher area under the receiver operating characteristic curve (AUC) than alternatively-trained classifiers. Our cost model accurately predicts the number of annotations necessary to obtain balanced classes. The accuracy of our prediction is verified by empirically-observed costs. Finally, we find that over-sampling the minority class yields a marginal improvement in classifier accuracy but the improved performance comes at the expense of greater annotation cost.

3.2 Introduction

In most supervised classification schemes, a training set of exemplars from each class is used to train a classifier to distinguish between the different object classes. The training exemplars (e.g. images, pixels, regions of interest) usually have a semantic label assigned to them by an expert describing a category of interest or class to which they belong. Each training exemplar serves as an observation of the domain space; as the space is sampled more completely, the resulting classifier should achieve greater accuracy when predicting class labels for new, unlabeled (unseen) data. Thus, typically, the larger the training set, the greater the accuracy of the resulting classifier [46]. In most cases, the training set of labeled data for each of the object categories is generated by a human expert who manually annotates a pool of unlabeled samples by assigning a label to each exemplar.

The use of computers in histopathology analysis, known as digital pathology, is an increasingly common practice that promises to facilitate the detection, diagnosis, and treatment of disease [47]. Supervised classifiers have been applied in this context for a number of problems such as cancer detection and grading [8, 48, 6, 49, 50, 19]. If the objective of the classifier is to distinguish normal from cancerous regions of tissue, exemplars corresponding to each class need to be manually labeled by a domain expert (typically a pathologist). Figure 3.1 shows an image from such an annotation task, where a prostate tissue sample stained with hematoxylin and eosin (H&E) has been digitized at 40x optical magnification using a whole-slide scanner. In this case,

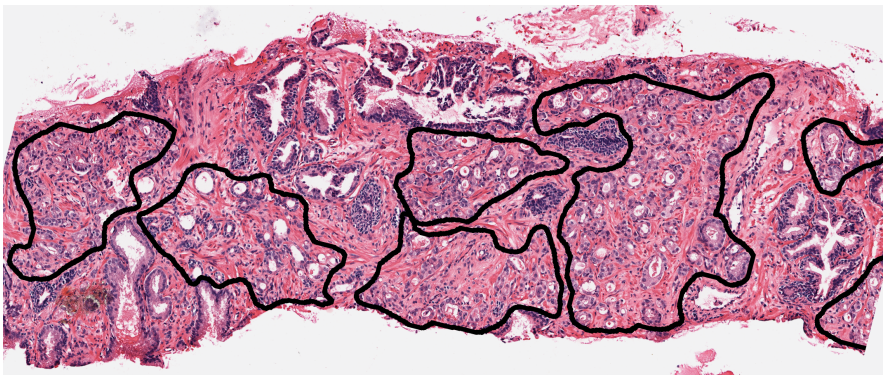


Figure 3.1: Annotation of CaP (black contour) on digital histopathology. CaP tissue often appears near and around non-CaP tissue, making annotation difficult and time-consuming.

the goal of the supervised classifier is to identify regions of carcinoma of the prostate (CaP, the target class). The black contour in Figure 3.1 indicates the target class and was placed manually by an expert pathologist. We have previously shown [8] that a supervised classifier can accurately distinguish between CaP and non-CaP, but the annotation process required to build a large training set is laborious, time consuming, and expensive. The digitized images can be over 2 gigabytes (several million pixels) in size, making it difficult to quickly identify cancerous regions within the digital slide. In addition, CaP often appears within and around non-CaP areas, and the boundary between these regions is not always clear (even to a trained expert). These factors increase the time, effort, and overall cost associated with training a supervised classifier in the context of digital pathology. To reduce the cost and effort involved in training these classifiers, it is important to utilize an intelligent labeling strategy.

Active learning (AL) is a method of classifier training wherein only "informative" exemplars are chosen for annotation from a pool of unlabeled samples. Informative samples are those which, if annotated and added to the training set, would increase the accuracy of the resulting trained classifier. This is in contrast to random learning (RL), wherein exemplars are annotated at random. With an RL-based training strategy, it is possible that many non-informative samples (samples that will not have a positive impact on classifier performance) will be annotated; clearly a wasted effort. Figure 3.2 illustrates the differences between RL (top row) and AL (bottom row). Several AL algorithms have been proposed [51, 9, 52, 53, 54, 55] to determine whether an unlabeled

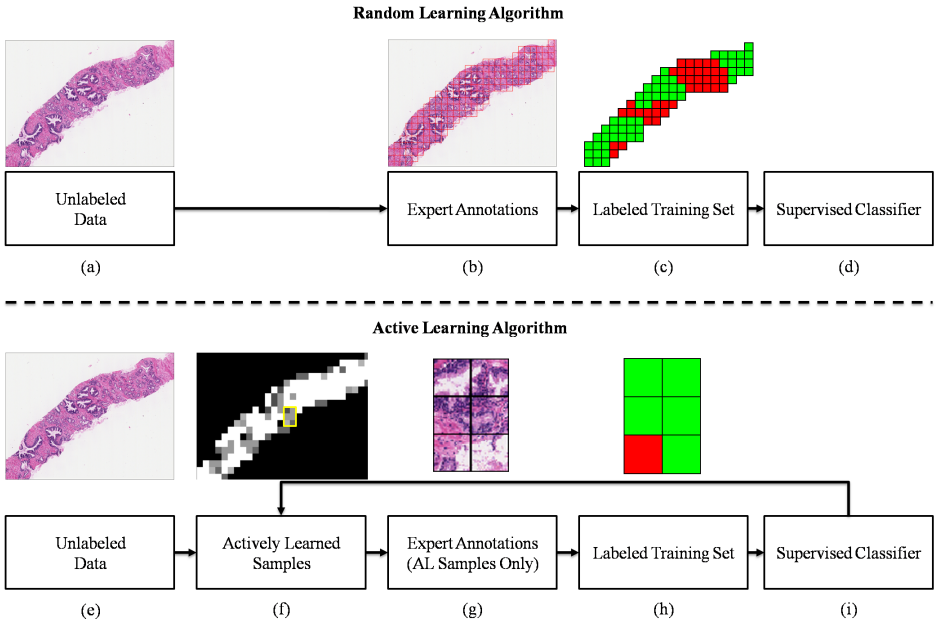


Figure 3.2: Comparison of Random Learning (RL, top row) and Active Learning (AL, bottom row) training processes. In RL, unlabeled data (a) is sent to an expert (b), who assigns a label to each sample in the image (c): red regions indicate cancer, and green indicates non-cancer. These labeled samples are used to train a supervised classifier (d). In AL, unlabeled samples (e) are analyzed to find informative samples (f), and only informative samples (g) are annotated for training (h). The supervised classifier (i) can be re-trained and used to identify new samples that may be informative. In the AL setup, only new samples that will improve classification accuracy are added.

sample is informative. Lee, et al. [56] and Veeramachaneni, et al. [57] used AL training techniques in the context of biomarker data.

Another major issue in supervised training involves the minority class problem, wherein the target class is under-represented in the dataset, relative to the non-target classes. A labeled training set comprises two sets of samples: $\mathbf{S}^{\mathrm{tr}}_{\omega_1}$ representing training samples from the target (minority) class, and $\mathbf{S}^{\mathrm{tr}}_{\omega_2}$ being the samples from the non-target (majority) class. In the minority class problem, $|\mathbf{S}^{\mathrm{tr}}_{\omega_1}| << |\mathbf{S}^{\mathrm{tr}}_{\omega_2}|$, where $|\cdot|$ indicates set cardinality. Several researchers [58, 59, 60, 61, 62] have shown that this training set will likely yield a classifier with lower accuracy and area under the receiver operating characteristic curve (AUC) compared with training sets where $|\mathbf{S}^{\mathrm{tr}}_{\omega_1}| = |\mathbf{S}^{\mathrm{tr}}_{\omega_2}|$ or $|\mathbf{S}^{\mathrm{tr}}_{\omega_1}| >> |\mathbf{S}^{\mathrm{tr}}_{\omega_2}|$. Weiss and Provost [58] showed that for several datasets, varying the percentage of the minority class in the training set alters the accuracy and AUC of the resulting classifiers, and that the optimal class ratio was found to be significantly

different from the "natural" ratio. Japkowicz and Stephen [59] found that the effect of the minority class problem depends on a number of factors, including the complexity of the target class and the size of the class disparity. Chawla, et al. [60] proposed mitigating the problem by over-sampling the minority class using synthetic samples; however, this method may simply introduce noise if the target class is too complex.

While some research has addressed the minority class problem in biomedical data [56, 63], there has been little related work in the realm of digital pathology. Cosatto, et al. [64] applied a support-vector machine AL method [54] in training a classifier for grading nuclear pleomorphism on breast tissue histology, while Begelman, et al. [65] employed an AL-trained support vector machine classifier in building a telepathology system for prostate tissue analysis. However, these studies did not account for the minority class problem in the training set, which is particularly relevant in the context of digital pathology since the target class (cancer) is often observed far less often than the non-target class (non-cancer) and occupies only a small percentage of the overall tissue area. Ideally, an intelligent training strategy for this domain would combine AL while simultaneously addressing the minority class problem by maintaining a user-defined class ratio (class balancing). Zhu and Hovey [61] combined an entropy-based AL technique with over- and under-sampling to overcome the minority class problem for text classification, and found that over-sampling the minority class yielded the highest classifier performance. However, they did not investigate different class ratios and did not discuss the increased cost of the sampling techniques. Bloodgood and Vijay-Shanker [66] focused on an AL and classification method based on support-vector machines for unbalanced text and protein expression data; their approach involves estimating the class balance in the entire dataset, and then selecting samples to overcome this bias (as opposed to overcoming bias in the growing training set generated by AL).

While additional sampling can help to mitigate the minority class problem, this process requires more annotations compared to a training set with unbalanced classes. Because the cost of obtaining each annotation is high, we must have a way of predicting how many annotations will be required to obtain a class-balanced training set of a pre-defined size. These predictions are critical for determining, *a priori*, the amount of

resources (time, money, manpower) that will be employed in developing a supervised classifier. An analytical cost model will enable us to predict the cost involved in training the supervised classifier. Additionally, such a model will provide some insight into the relationship between (1) the size of a training set, (2) its class balance, and (3) the number of annotations required to achieve a predefined target accuracy.

3.3 Contributions and Significance

In this work, we develop an Active Learning (AL) based classifier training strategy that also accounts for the minority class problem. This training strategy is referred to as "Class-Balanced Active Learning" (CBAL). We apply CBAL to the problem of building a supervised classifier to distinguish between CaP and non-CaP on images of prostate histopathology. In this domain, training samples are difficult and expensive to obtain, and the target class (CaP) is rare in relation to the non-target class; thus, we expect CBAL to yield large benefits in terms of training cost. Our mathematical model is used to predict the cost of building a training set of pre-defined size and class ratio. This is, to the best of our knowledge, the first in-depth investigation and modeling of AL-based training for supervised classifiers that also specifically addresses the minority class problem in the context of digital pathology. However, CBAL training can be easily applied to other domains where obtaining annotated training samples is a time-consuming and difficult task, and where the target and non-target class ratios are not balanced.

The rest of the paper is organized as follows. In Section 3.4 we describe the theory behind CBAL, followed by a description of the algorithms and model implementation in Section 3.5. In Section 3.6 we describe our experimental design, and in Section 3.7 we present the results and discussion. Concluding remarks are presented in Section 3.8.

Symbol	Description	Symbol	Description
$r \in R$	Dataset of image patches	t	Iteration of *ActiveLearn*
$\mathbf{S}^{\text{tr}}$, $\mathbf{S}^{\text{te}}$	Unlabeled training, testing pools	Φ	Training methodology
$\mathbf{S}_t^{\text{E}}$,$\widehat{\mathbf{S}}_t^{\text{E}}$	Eligible, annotated samples	$S_{t,\Phi}^{\text{tr}}$	Samples labeled via Φ at t
$\mathcal{T}_t$	Fuzzy classifier using $S_{t,\Phi}^{\text{tr}}$	$k_{1,t}$, $k_{2,t}$	Number of $\mathbf{S}_t^{\text{E}}$ samples from ω_1, ω_2
M	Number of votes to generate $\mathcal{T}_t$	ω_1, ω_2	Possible classes of r
τ	Confidence margin	$r \hookrightarrow \omega_1$	Membership of r in class ω_1
θ	Classifier threshold for $\mathcal{T}_t$	$\widehat{k_1}$, $\widehat{k_2}$	Number of $\widehat{\mathbf{S}}_t^{\text{E}}$ samples from ω_1, ω_2
$p_t(r \hookrightarrow \omega_1)$	Probability of observing $r \hookrightarrow \omega_1$	N_t	Samples added to training set at t
P_Δ	Model confidence	$\widehat{P_t}$	Probability of observing $\widehat{k_1}$ samples
$\mathcal{A}_t$	Accuracy of trained classifier at t	$\mathcal{L}$	Training cost after T iterations

Table 3.1: List of the commonly used notation and symbols.

3.4 Modeling the Annotation Cost of Class Balancing in Training

3.4.1 Notation and Symbols

A table containing commonly used notation and symbols is presented in Table 3.1. Our data comprises a set of square image regions $r \in R$ on digitized prostate images, represented by the red squares in Figure 3.2 (e). The regions $r \in R$ are divided into an unlabeled training pool, $\mathbf{S}^{\text{tr}}$, and an independent labeled testing pool, $\mathbf{S}^{\text{te}}$. Each sample has been identified as either belonging to the minority class ω_1 (in this case the cancer class) or the majority (non-cancer) class, ω_2. We denote membership of sample $r \in R$ in the minority class ω_1 as $r \hookrightarrow \omega_1$, and these samples are "minority class samples." At iteration $t \in \{0, 1, \cdots, T\}$ of AL, the labeled training set is denoted as $S_{t,\Phi}^{\text{tr}}$, where Φ denotes the training methodology and T is the maximum number of iterations. At each iteration t, a set of M weak binary classifiers is trained by $S_{t,\Phi}^{\text{tr}}$ and used to build a strong classifier, $\mathcal{T}_t(r) \in \{0, \cdots, 1\}$. The selectivity of the AL algorithm is parameterized by $\tau \in \{0, \cdots, 1\}$, the confidence margin. We denote by $\widehat{k_1}$ and $\widehat{k_2}$ the desired number of samples $r \in R$ in the final training set for which $r \hookrightarrow \omega_1$ and $r \hookrightarrow \omega_2$, respectively. The total number of samples annotated at any iteration t is denoted as N_t.

3.4.2 Theory of CBAL

Definition 1. *The set of informative samples (eligible for annotation),* $\mathbf{S}_t^E$*, at any iteration* t *is given by the set of samples* $r \in R$ *for which* $0.5 - \tau \leq \mathcal{T}_t(r) \leq 0.5 + \tau$.

The value of $\mathcal{T}_t(r)$ denotes the classification confidence, where $\mathcal{T}_t(r) = 1$ indicates strong confidence that $r \hookrightarrow \omega_1$, and $\mathcal{T}_t(r) = 0$ indicates confidence that $r \hookrightarrow \omega_2$. The number of samples $r \in \mathbf{S}_t^{\mathrm{E}}$ for which $r \hookrightarrow \omega_1$ and $r \hookrightarrow \omega_2$ are denoted $k_{1,t}$ and $k_{2,t}$, respectively. The likelihood of randomly selecting a sample $r \hookrightarrow \omega_1$ from $\mathbf{S}_t^{\mathrm{E}}$ is $p_t(r \hookrightarrow \omega_1) = \frac{k_{1,t}}{k_{1,t}+k_{2,t}}$. The number annotated in class ω_2 is $N_t - \widehat{k_1}$.

Proposition 1. *Given the probability $p_t(r \hookrightarrow \omega_1)$ of observing a sample $r \hookrightarrow \omega_1$ at any iteration t, the probability $\widehat{P}_t$ of observing $\widehat{k_1}$ samples from class ω_1 after annotating N_t samples is:*

$$\widehat{P}_t = \sum_{\alpha=0}^{\binom{N_t}{\widehat{k_1}}} \left[p_t(r \hookrightarrow \omega_1)\right]^{\widehat{k_1}} \left[1 - p_t(r \hookrightarrow \omega_1)\right]^{N_t - \widehat{k_1}}. \tag{3.1}$$

Proof. Revealing the label of a sample $r \in \mathbf{S}_t^E$ is an independent event with two possible outcomes: observation of class ω_1 or ω_2. The probability of success (i.e. observing a minority class sample) is $p_t(r \hookrightarrow \omega_1)$, and the probability of failure is $p_t(r \hookrightarrow \omega_2) = 1 - p_t(r \hookrightarrow \omega_1)$ in the two class case. Therefore, we can describe the probability of $\widehat{k_1}$ successes after N_t observations by the binomial distribution:

$$P_t = \binom{N_t}{\widehat{k_1}} \left[p_t(r \hookrightarrow \omega_1)\right]^{\widehat{k_1}} \left[1 - p_t(r \hookrightarrow \omega_1)\right]^{N_t - \widehat{k_1}}. \tag{3.2}$$

However, we wish to minimize N_t, as the annotation process stops as soon as $\widehat{k_1}$ successes are reached, while Equation 3.2 will force no less than a specified N_t annotations to take place. Therefore as N_t increases, we calculate the sum of probabilities from 0 to $\binom{N_t}{\widehat{k_1}}$ indicating the cumulative probability of $\widehat{k_1}$ successes in N_t trials with the last trial being the final success. □ □

The consequence of Proposition 1 is that as N_t (i.e. the training cost in annotations) increases, $\widehat{P}_t$ also increases, indicating a greater likelihood of observing $\widehat{k_1}$ samples $r \hookrightarrow \omega_1$. We denote as P_Δ the target probability for the model to represent the degree of certainty that, within N_t annotations, we have achieved our $\widehat{k_1}$ samples $r \in R$ for which $r \hookrightarrow \omega_1$.

Proposition 2. *Given a target probability P_Δ, the number of annotations required*

before $\widehat{k_1}$ *minority class samples are observed in* $\mathbf{S}^E$ *is:*

$$N_t = \underset{\widehat{k_1} \leq x \leq |\mathbf{S}^{tr}|}{\operatorname{argmin}} \left[P_\Delta - \sum_{\alpha=0}^{\binom{x}{\widehat{k_1}}} [p_t(r \hookrightarrow \omega_1)]^{\widehat{k_1}} [1 - p_t(r \hookrightarrow \omega_1)]^{x-\widehat{k_1}} \right]. \tag{3.3}$$

Proof. We wish to find the value of N_t that causes Equation 1 to match our target probability, P_Δ. When that happens, $\widehat{P}_t = P_\Delta$ and $\widehat{P}_t - P_\Delta = 0$. Using a minimization strategy, we obtain the value of N_t. □ □

Proposition 2 gives us an analytical formulation for N_t. Note that Equation 3 returns the smallest N_t that matches the P_Δ. The possible values of N_t range from $\widehat{k_1}$, in which case exactly $N_t = \widehat{k_1}$ annotations are required, to $N_t = |\mathbf{S}^{\text{tr}}|$, in which case the entire dataset is annotated before obtaining $\widehat{k_1}$ samples. Note that we are assuming that there are at least $\widehat{k_1}$ samples in the unlabeled training set from which we are sampling.

3.5 Algorithms and Implementation

3.5.1 AL Algorithm for Selecting Informative Samples

The CBAL training strategy consists of two algorithms: *ActiveTrainingStrategy*, for selecting informative samples, and *MinClassQuery*, for maintaining class balance. Algorithm *ActiveTrainingStrategy*, detailed below, requires a pool of unlabeled samples, $\mathbf{S}^{\text{tr}}$, from which samples will be drawn for annotation, as well as a parameter for maximum iterations T. This parameter can be chosen according to the available training budget or through a pre-defined stopping criterion. The output of the algorithm will be a fully annotated training set $S^{\text{tr}}_{T,\Phi}$ as well as the classifier trained using training set $\mathcal{T}_T$.

The identification of the informative samples occurs in Step 3, wherein a fuzzy classifier $\mathcal{T}_t$ is generated from a set of M weak binary decision trees [45] that are combined via bagging [67]. Informative samples are those samples for which half of the M weak binary decision trees disagree; that is, samples for which $\mathcal{T}_t(r) = \frac{1}{2} \pm \tau$. This approach is similar to the Query-by-Committee (QBC) AL algorithm [51, 9]. While several alternative algorithms are available [52, 53, 54, 55], the QBC algorithm is used here because of its intuitive description of sample informativeness.

Algorithm ActiveTrainingStrategy()
Input: $\mathbf{S}^{\text{tr}}$, T
Output: $S^{\text{tr}}_{T,\Phi}$, $\mathcal{T}_T$
begin
0. initialization: create bootstrap training set $S^{\text{tr}}_{0,\Phi}$, set $t = 0$
1. **while** $t < T$
2. Create classifier $\mathcal{T}_t$ from training set $S^{\text{tr}}_{t,\Phi}$;
3. Find eligible sample set $\mathbf{S}^{\text{E}}_t$ where $\mathcal{T}_t(r) = \frac{1}{2} \pm \tau$;
4. Annotate K eligible samples via *MinClassQuery()* to obtain $\widehat{\mathbf{S}}^{\text{E}}_t$;
5. Remove $\widehat{\mathbf{S}}^{\text{E}}_t$ from $\mathbf{S}^{\text{tr}}$ and add to $S^{\text{tr}}_{t+1,\Phi}$;
6. $t = t + 1$;
7. **endwhile**
8. **return** $\mathcal{T}_T$, $S^{\text{tr}}_{T,\Phi}$;
end

3.5.2 Obtaining Annotations While Maintaining Class Balance

Algorithm *MinClassQuery* is used by *ActiveTrainingStrategy* to select samples from the set of eligible samples, $\mathbf{S}^{\text{E}}_t$, according to a class ratio specified by $\widehat{k_1}$ and $\widehat{k_2}$. Recall that $K = \widehat{k_1} + \widehat{k_2}$, and so $K > 0$.

We expect that there will be many more samples from ω_2 (the majority class) than from ω_1. Because these samples are being annotated, they are removed from the unlabeled eligible sample pool $\mathbf{S}^{\text{E}}_t$ in Step 7; however, since the resources have been expended to annotate them, they can be saved for future iterations.

3.5.3 Updating Cost Model and Stopping Criterion Formulation

At each iteration, we can calculate N_t using equation 3.1. We can estimate $p_0(r \hookrightarrow \omega_1)$ based on the size of the target class observed empirically from the initial training set ($< 10\%$); for $t > 0$, we update the probability of observing a minority class sample using the following equation:

$$p_{t+1}(r \hookrightarrow \omega_1) = \frac{k_{1,t} - \widehat{k_1}}{k_{1,t} + k_{2,t} - N_t}, \tag{3.4}$$

and N_{t+1} is re-calculated via the minimization of Equation 3.3. If $\{r \in \mathbf{S}^{\text{tr}} | r \hookrightarrow \omega_1\} = \emptyset$, then $k_{1,t} - \widehat{k_1} = 0$ and thus $p_{t+1}(r \hookrightarrow \omega_1) = 0$. If there are no remaining samples in $\mathbf{S}^{\text{tr}}$,

Algorithm MinClassQuery()
Input: $\mathbf{S}_t^{\mathrm{E}}$, $K > 0$, $\widehat{k_1}$, $\widehat{k_2}$
Output: $\widehat{\mathbf{S}}_t^{\mathrm{E}}$
begin
0. initialization: $\widehat{\mathbf{S}}_t^{\mathrm{E}} = \emptyset$, $k_1' = 0$, $k_2' = 0$
1. **while** $|\widehat{\mathbf{S}}_t^{\mathrm{E}}| \neq K$
2. Find class ω_i of a random sample $r \in \mathbf{S}_t^{\mathrm{E}}$, $i \in \{1, 2\}$;
3. **if**$k_i' < \widehat{k_i}$
4. Remove r from $\mathbf{S}_t^{\mathrm{E}}$ and add to $\widehat{\mathbf{S}}_t^{\mathrm{E}}$;
5. $k_i' = k_i' + 1$;
6. **else**
7. Remove r from $\mathbf{S}_t^{\mathrm{E}}$;
8. **endif**
9. **return** $\widehat{\mathbf{S}}_t^{\mathrm{E}}$;
end

then $k_{1,t} + k_{2,t} = N_t$ and $p_{t+1}(r \hookrightarrow \omega_1)$ is undefined. Essentially we must assume that (1) there are at least some samples $r \in \mathbf{S}^{\mathrm{tr}}$ for which $r \hookrightarrow \omega_1$, and (2) $\mathbf{S}^{\mathrm{tr}} \neq \emptyset$. The cost of the entire training is calculated by summing N_t for all t:

$$\mathcal{L} = \sum_{t=1}^{T} N_t. \tag{3.5}$$

ActiveTrainingStrategy repeats until one of two conditions is met: (1) $\mathbf{S}^{\mathrm{tr}}$ is empty, or (2) the maximum number of iterations T is reached. A stopping criterion can be trained offline to determine the value of T as the smallest t that satisfies:

$$|\mathcal{A}_t - \mathcal{A}_{t-1}| \leq \delta, \tag{3.6}$$

where δ is a similarity threshold and $\mathcal{A}_t$ is the accuracy of classifier $\mathcal{T}_t$ (as evaluated on a holdout training set). Thus, when additional training samples no longer increase the resulting classifier's accuracy, the training can cease. An assumption in using this stopping criterion is that adding samples to the training set will not *decrease* classifier accuracy. The total number of iterations T corresponds to the size of the final training set and can be specified manually or found using a stopping criterion discussed below. Classifiers that require a large training set will require a large value for T, increasing cost.

3.6 Experimental Design

3.6.1 Data Description

We apply the CBAL training methodology to the problem of prostate cancer detection from biopsy samples. This is a similar task to our previous work [8], wherein we discriminated between cancerous and non-cancerous regions in a pixel-wise fashion. Glass slides containing prostate biopsy samples are digitized at 40x magnification and are divided into sets of square regions, $r \in R$. Regions are 30-by-30 square pixel regions; this size was empirically determined as the size required to distinguish cancer attributes from non-cancer. Ground truth annotation for cancer regions is performed manually by an expert pathologist. A total of 100 images were analyzed from 58 patients yielding over 12,000 image regions.

3.6.2 Feature Extraction

In [8] we identified several hundred textural features useful for discriminating between cancerous and non-cancerous regions, at pixel-level resolution. The 14 most discriminating texture features were selected that were found to best distinguish between cancerous and benign tissue. The feature set comprises three different classes of texture descriptors; examples of these feature types are given in Figure 3.3. These pixel-wise features are calculated for each 30-by-30 region. Each region r was then represented by the average value of the feature calculated over all pixels.

First-order Statistical Features

First-order features are statistics calculated directly from the pixel values in the image. These include the mean, median, and standard deviation of the pixels within a window size, as well as Sobel filters and directional gradients. Of these features, two were considered highly discriminating: The standard deviation and the range of pixel intensities.

Second-order Co-occurrence Features

Co-occurrence image features are based on the adjacency of pixel values in an image. An adjacency matrix is created where the value of the ith row and the jth column equals the number of times pixel values i and j appear within a fixed distance of one another. A total of thirteen Haralick texture features [29] are calculated from this co-adjacency matrix, of which 5 were found to be highly discriminating: information measure, correlation, energy, contrast variance, and entropy.

Steerable Filter Features

To quantify spatial and directional textures in the image, we utilize a steerable Gabor filter bank [31]. The Gabor filter is parameterized by frequency and orientation (phase) components; when convolved with an image, the filter provides a high response for textures that match these components. We compute a total of 40 filter banks, of which 7 were found to be informative, from a variety of frequency and orientation values.

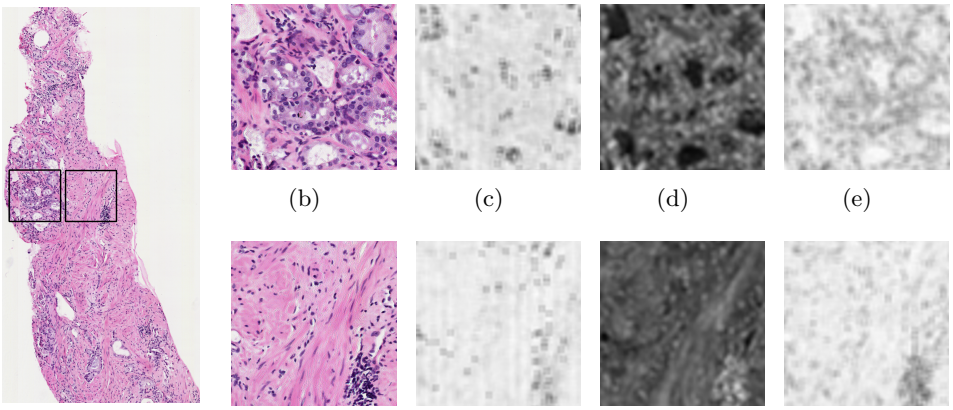


Figure 3.3: Examples of the feature types extracted on two ROIs from a biopsy sample (a), identified by black squares. Shown are (b), (f) the original tissue image, (c), (g) a greylevel texture image (standard deviation value), (d), (h) a Haralick texture image (entropy of the co-adjacency matrix), and (e), (e) a Gabor filter feature image. The top row (b)-(e) indicates a cancerous region, while the bottom row (f)-(i) is a benign region.

3.6.3 Evaluation of Training Set Performance via Probabilistic Boosting Trees

Evaluation of $S_{t,\Phi}^{\text{tr}}$ is done by testing the trained classifier's accuracy. To avoid biasing the results, we wish to use a different classifier than $\mathcal{T}_t$ for evaluation; a probabilistic boosting tree (PBT) [68], denoted $\mathcal{T}_t'$, is employed. The PBT combines AdaBoost [9] and decision trees [45] and recursively generates a decision tree where each node is boosted with M weak classifiers. The classifier output, $\mathcal{T}_t'(r)$, is the probability that sample r belongs to the target class. The PBT is used to classify an independent testing set $\mathbf{S}^{\text{te}}$ (where $\mathbf{S}^{\text{te}} \cap \mathbf{S}^{\text{tr}} = \emptyset$) via area under the receiver operating characteristic curve (AUC) and classifier accuracy. The hard classification for $r \in \mathbf{S}^{\text{te}}$ is denoted as:

$$\tilde{\mathcal{T}}_t(r) = \begin{cases} 1 & \text{if } \mathcal{T}_t'(r) > \theta \\ 0 & \text{otherwise,} \end{cases} \tag{3.7}$$

where θ is a classifier-dependent threshold. For region r, the ground truth label is denoted as $\mathcal{G}(r) \in \{0, 1\}$, where a value of 1 indicates class ω_1 and 0 indicates class ω_2. The resulting accuracy at iteration t is denoted as:

$$\mathcal{A}_t = \frac{1}{|R|} \sum_r \begin{cases} 1 & \text{if } \mathcal{G}(r) = \tilde{\mathcal{T}}_t(r) \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

We generate receiver operating characteristic (ROC) curves by calculating the classifier's sensitivity and specificity at various decision thresholds $\theta \in \{0, \cdots, 1\}$. Each value of θ yields a single point on the ROC curve, and the area under the curve (AUC) measures the discrimination between cancer and non-cancer regions. The accuracy can then be calculated by setting θ to the operating point of the ROC curve.

3.6.4 List of Experiments

We perform three sets of experiments to analyze different facets of the active learning training methodology.

Experiment 1: Comparison of CBAL performance with Alternate Training Strategies

We compare the performance of CBAL with four alternative training strategies to show that CBAL training will yield a classifier with greater performance versus a training set of the same size trained using an alternative method.

- Unbalanced Active Learning (UBAL): The class ratio is not controlled; eligible samples $\mathbf{S}_t^{\mathrm{E}}$ determined via AL are randomly annotated and added to $\widehat{\mathbf{S}}_t^{\mathrm{E}}$.
- Class Balanced Random Learning (CBRL): All unlabeled samples in $\mathbf{S}^{\mathrm{tr}}$ are eligible for annotation, while holding class balance constant as described in *MinClassQuery*.
- Unbalanced Random Learning (UBRL): All unlabeled samples are queried randomly. This is the classic training scenario, wherein neither class ratio nor informative samples are explicitly controlled.
- Full Training (FULL): All available training samples are used. This represents the performance when the entire training set is annotated and available (an ideal scenario).

In random learning (RL), all samples in the unlabeled pool $\mathbf{S}^{\mathrm{tr}}$ are "eligible" for annotation; that is, $\mathbf{S}^{\mathrm{E}} = \mathbf{S}^{\mathrm{tr}}$. In unbalanced class experiments, the *MinClassQuery* algorithm is replaced by simply annotating K random samples (ignoring class) and adding them to $\widehat{\mathbf{S}}^{\mathrm{E}}$. The FULL training strategy represents the scenario when all possible training data is used.

The classifier is tested against the independent testing pool, $\mathbf{S}^{\mathrm{te}}$. In these experiments, $T = 40$, the confidence margin was $\tau = 0.5$, and the number of samples added at each iteration was $K = 2$. In the balanced experiments, $\widehat{k_1} = \widehat{k_2} = 1$. A total of 12,588 image regions were used in the overall dataset; 1,346 were randomly selected for $\mathbf{S}^{\mathrm{te}}$, and 11,242 for $\mathbf{S}^{\mathrm{tr}}$. The true ratio of non-cancer to cancer regions in $\mathbf{S}^{\mathrm{tr}}$ was approximately 25:1 (4% belonged to the cancer class). A total of 10 trials were performed, randomly selecting $\mathbf{S}^{\mathrm{tr}}$ and $\mathbf{S}^{\mathrm{te}}$ each time.

Experiment 2: Effect of Training Set Class Ratio on Accuracy of Resulting Classifier

To explore the effect of training set class ratio on the performance of the resulting classifier, the CBAL methodology was used, setting $K = 10$ and varying $\widehat{k_1}$ and $\widehat{k_2}$ such that the percentages of the training set consisting of minority samples vary from 20% ($\widehat{k_1} = 2$, $\widehat{k_2} = 8$) to 80% ($\widehat{k_1} = 8$, $\widehat{k_2} = 2$). Each set of parameters was used to build a training set, which in turn was used to build a classifier that was evaluated on the same independent testing set $\mathbf{S}^{\text{te}}$.

Experiment 3: Comparison of Cost Model Predictions with Empirical Observations

At each step of the AL algorithm, we estimate N_t for obtaining balanced classes as described in Section 3.4. The goal of this experiment was to empirically evaluate whether our mathematical model could accurately predict the cost of obtaining balanced classes at each iteration, and could thus be used to predict the cost of classifier training for any problem domain. For these calculations, we set the initial class probability $p_0(\omega_1) = 0.04$, based on the observations of the labeled data used at the beginning of the AL process. Additionally, we set the desired sample numbers to correspond with the different class ratios listed in Experiment 2, from 20% minority class samples ($\widehat{k_1} = 2$, $\widehat{k_2} = 8$) to 80% ($\widehat{k_1} = 8$, $\widehat{k_2} = 2$). The aim of this experiment was to investigate the relationship between the cost of a specific class ratio and the performance of $\mathcal{T}_T'$.

3.7 Results and Discussion

3.7.1 Experiment 1: Comparison of CBAL performance with Alternate Training Strategies

Examples of confidence or likelihood scenes generated by $\mathcal{T}_T'$ are shown in Figure 3.4, obtained at iteration $T = 40$ (since $K = 2$, these images represent the classifier's performance using 80 total samples). Figures 3.4 (a), (d) show images with benign regions marked in red boundaries and cancerous regions in black. Figures 3.4 (b), (e)

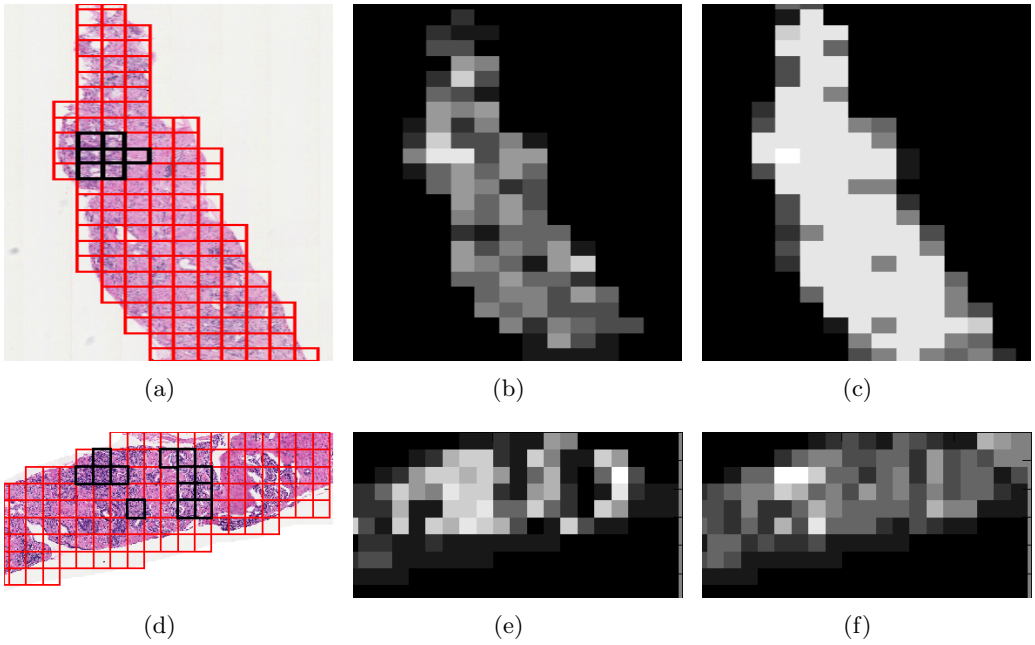
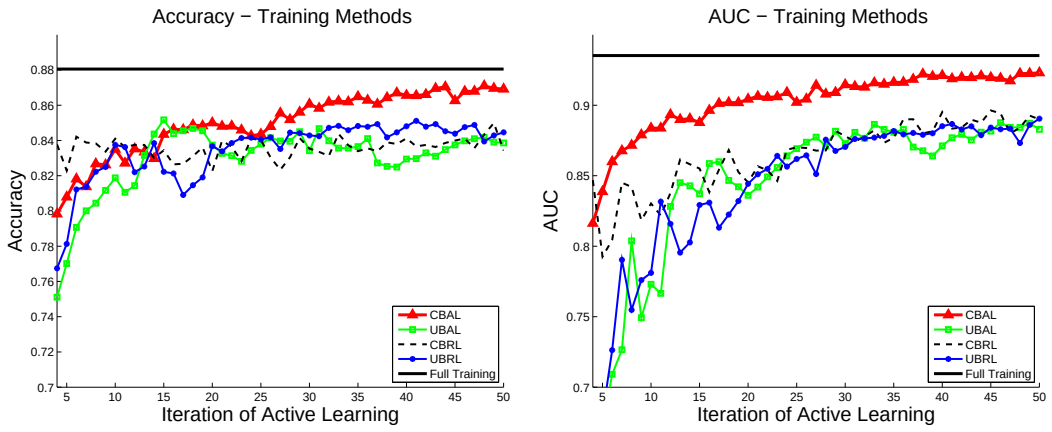


Figure 3.4: Qualitative results of the final PBT classifier $\mathcal{T}'_T$. Shown in (a), (d) are the segmented cancer region, (b), (e) the probability scene obtained through the CBAL classifier, and (c), (f) the probability scene obtained via CBRL. The intensity of a region is determined by $\mathcal{T}'_T(r)$.

show the confidence scenes obtained via the CBAL training strategy, and (c), (f) are obtained via CBRL training. High intensity regions represent high classifier confidence that $r \hookrightarrow \omega_1$, while dark regions indicate confidence that $r \hookrightarrow \omega_2$. In both cases, the CBRL training fails to properly find the cancer regions, either returning large numbers of false positives (Figure 3.4 (c)) or failing to fully identify the cancer area (Figure 3.4 (f)). This difference (high false positives in one case, high false negatives in another) is most likely due to the inability of random learning to accurately define the classes, given the small training set size. Thus, given the constraints on training set size, a CBAL-trained classifier can out-perform a randomly-trained classifier.

Quantitative classification results are plotted in Figure 3.5 as accuracy (Figure 3.5 (a)) and area under the ROC curve (Figure 3.5 (b)) as a function of the number of training samples in the set S_t^{tr} for $1 \leq t \leq 40$. In each plot, the FULL training set corresponds to the straight black line, CBAL is the red triangle line, CBRL is a black dashed line, UBAL is a green squared line, and UBRL is a blue circled line. Note that the FULL line indicates the maximum achievable classifier accuracy for a given



(a) (b)

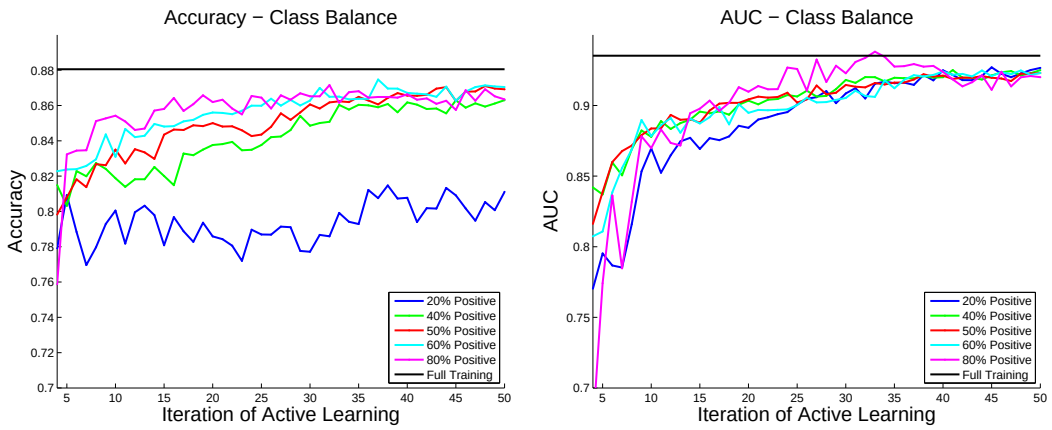
Figure 3.5: Quantitative results of the classifier, $\mathcal{T}_t'$, for $t \in \{1, 2, \cdots, T\}$. Shown are (a) accuracy and (b) AUC values for the PBT classifier, evaluated at each iteration.

training set; thus, the closer a training set gets to the straight black line, the closer it is to optimal performance.

The AUC values for CBAL approach the FULL training with 60 samples ($t = 30$) while CBRL, UBRL, and UBAL have approximately 0.05 lower AUC at those sample sizes. Accuracy for CBAL remains similar to other methods until $t = 30$, at which point CBAL out-performs other methods by approximately 3%. CBRL, UBRL, and UBAL do not perform as well as CBAL for the majority of our experiments, requiring a larger number of samples to match the accuracy and AUC of CBAL.

3.7.2 Experiment 2: Effect of Training Set Class Ratio on Accuracy of Resulting Classifier

Figure 3.6 shows the effects of varying training class ratios on the resulting classifier's performance for the prostate cancer detection problem. Shown is the performance of the PBT classifier at each iteration of the AL algorithm using 20% minority samples (blue line), 40% (green line), 50% (red line), 60% (cyan line), and 80% (magenta line), for both accuracy (Figure 3.6 (a)) and AUC (Figure 3.6 (b)). The AUC curves are similar for all class ratios, although the training set that uses 80% minority class samples tends to perform slightly better. Thus, by over-representing the minority class, we achieve greater performance in terms of accuracy. Noted that while changing the class ratio



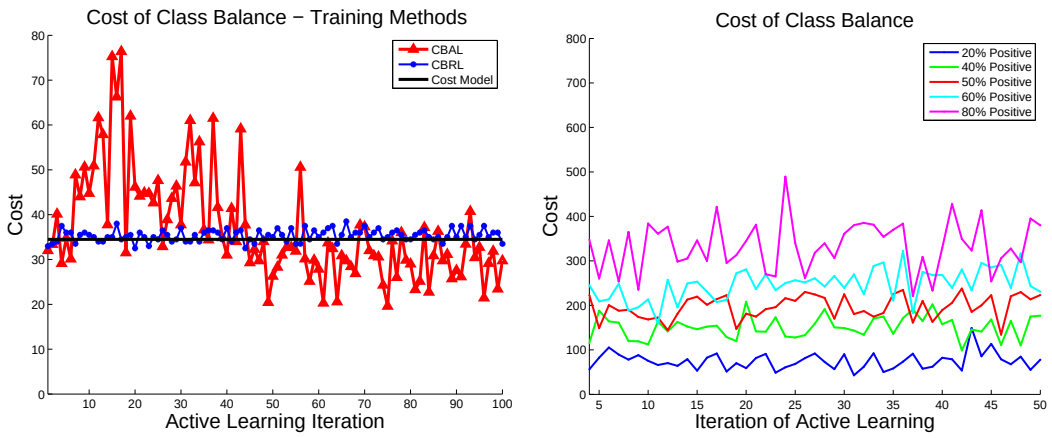
(a) (b)

Figure 3.6: Performance of the PBT classifier trained using training sets with different percentages of samples for which $r \hookrightarrow \omega_1$. Shown are the (a) accuracy and (b) AUC values for the trained classifier at each iteration, using $p_0(r \hookrightarrow \omega_1) = 0.04$.

had different effects on accuracy and AUC a similar trend was reported by Weiss and Provost [58], who found that altering the class ratio of a training set for a classifier affected AUC and accuracy differently (although there was no specific trend across multiple datasets).

3.7.3 Experiment 3: Comparison of Cost Model Predictions with Empirical Observations

Figure 3.7 (a) shows the results of cost modeling simulations. The predicted cost, found by solving for N_t in Equation 3.1, is plotted as a function of t (solid black line) with $p_0(r \hookrightarrow \omega_1) = 0.04$ along with the empirically observed costs of CBRL (blue dotted line) and CBAL (red triangle line) with $\widehat{k_1} = \widehat{k_2} = 5$. At each t, the plots show how many annotations were required before class balancing was achieved. We can see that the simulation predicts the number of annotations required to achieve class balance at each iteration within approximately 10-20 annotations. Additionally, we see that the empirically observed costs are greatly varied, particularly for $t < 50$; this is due to the fact that the number of annotations required to achieve class balance depends greatly on (1) the current training set, (2) the remaining samples in the unlabeled pool, and (3) the order in which eligible samples are chosen for annotation.



(a) (b)

Figure 3.7: (a) Plot of annotations N_t required for class balance as a function of t; shown are CBAL (blue line), CBRL (red dashed line), and the predicted N_t from Equation 3.1 (black line). (b) The cost of obtaining a specific class ratio as iterations increase. If a high percentage of minority class samples is desired, the cost increases.

While it may seem from Figure 3.6 that the strategy yielding best performance would be to over-sample the minority class as much as possible, we also plotted the empirical cost values N_t for each of the class ratios from Experiment 2 in Figure 3.7 (b). We find that as the percentage of the minority class increases, the cost associated with each iteration of the AL algorithm also increases. This is due to the fact that as the minority class is over-sampled, more annotations are required to find additional minority samples. While there is some increase in accuracy by over-sampling the dataset, the annotation cost increases by an order of magnitude. Thus, the optimal strategy will need to balance the increase in accuracy with the constraints of the overall annotation budget.

3.8 Concluding Remarks

In this work we present a strategy for training a supervised classifier when the costs of training are high, and where the minority class problem exists. Our strategy, Class-Balanced Active Learning (CBAL), has the following characteristics: (1) Active Learning (AL) is used to select informative samples for annotation, thus ensuring that each

annotation is highly likely to improve classifier performance. (2) Class ratios are specifically addressed in this training strategy to prevent the training set from being biased toward the majority class. (3) A mathematical model is used to predict the number of annotations required before the specified class balance is reached. We applied these techniques to the task of quantitatively analyzing digital prostate tissue samples for presence of cancer, where the CBAL training method yielded a classifier with accuracy and AUC values similar to those obtained with the full training set using fewer samples than the unbalanced AL, class-balanced random learning, or unbalanced random learning methods. Our mathematical cost model was able to predict the number of annotations required to build a class-balanced training set within 20 annotations, despite the large amount of variance in the empirically observed costs. This model is critical in determining, *a priori*, what the cost of training will be in terms of annotations, which in turn translates into the time and effort expended by the human expert in helping to build the supervised classifier. We found that by specifying class ratios for the training set that favor the minority class (i.e. over-sampling), the resulting classifier performance increased slightly; however, the cost model predicted a large increase in the cost of training, as a high percentage of minority class samples requires more annotations to build. Thus, an optimal training strategy must take into account the overall training budget and the desired accuracy.

Some of the specific findings in this work, such as the observation that over-representing the minority class yields a slightly higher classifier performance, may be specific to the dataset considered here. Additionally, the observation that the AL algorithm has a large amount of variance in the empirically-observed costs (particularly at the beginning of training) indicates that the eligible sample set is unpredictable with respect to class compositions. This behavior may not necessarily be duplicable with different datasets or AL strategies, both of which will yield eligible sample sets with different class compositions. However, by combining AL and class balancing, we have developed a general training strategy that should be applicable to most supervised classification problems where the dataset is expensive to obtain and which suffers from the minority class problem. These problems are particularly prevalent in medical image analysis

and digital pathology, where the costs of classifier training are very high and an intelligent training strategy can help save great amounts of time and money. Training is an essential and difficult part of supervised classification, but the integration of AL and intelligent choice of class ratios, as well as the application of a general cost model, will help researchers to plan the training process more quickly and effectively. Future work will involve extensions of our framework to the multi-class case, where relationships between multiple classes with different distributions must be taken into account.

Chapter 4

Consensus of Ambiguity: Theory and Application of Active Learning for Biomedical Image Analysis

4.1 Abstract

Supervised classifiers require manually labeled training samples to classify unlabeled objects. Active Learning (AL) can be used to selectively label only "ambiguous" samples, ensuring that each labeled sample is maximally informative. This is invaluable in applications where manual labeling is expensive, as in medical images where annotation of specific pathologies or anatomical structures is usually only possible by an expert physician. Existing AL methods use a single definition of ambiguity, but there can be significant variation among individual methods. In this paper we present a consensus of ambiguity (CoA) approach to AL, where only samples which are consistently labeled as ambiguous across multiple AL schemes are selected for annotation. CoA-based AL uses fewer samples than Random Learning (RL) while exploiting the variance between individual AL schemes to efficiently label training sets for classifier training. We use a consensus ratio to determine the variance between AL methods, and the CoA approach is used to train classifiers for three different medical image datasets: 100 prostate histopathology images, 18 prostate DCE-MRI patient studies, and 9,000 breast histopathology regions of interest from 2 patients. We use a Probabilistic Boosting Tree (PBT) to classify each dataset as either cancer or non-cancer (prostate), or high or low grade cancer (breast). Trained is done using CoA-based AL, and is evaluated in terms of accuracy and area under the receiver operating characteristic curve (AUC). CoA training yielded between 0.01-0.05% greater performance than RL for the same training set size; approximately 5-10 more samples were required for RL to match the performance of CoA, suggesting that CoA is a more efficient training strategy.

4.2 Introduction

4.2.1 Using Consensus Methods for Certainty and Ambiguity

Ensemble classification algorithms such as bagging, boosting [45], and random forests [36] rely on some concept of consensus among several "weak" classifiers to generate a single "strong" result. Consensus, in the context of ensemble learning, describes agreement among several classification algorithms. For example, given a data object $\mathbf{x} \in \mathbb{R}^N$ belonging to one of c classes, $\omega_1, \cdots, \omega_c$, we can construct L classifiers $\mathcal{C}_l(\mathbf{x})$, for $l \in \{1, 2, \cdots, L\}$. The probability that $\mathbf{x}$ belongs to class ω_j, for $j \in \{1, 2, \cdots, c\}$, according to classifier l is denoted $p_l(\omega_j|\mathbf{x})$. While several classifier ensemble strategies seek to combine the weak learners using different rules, the underlying spirit of these methods is to assign the sample to the class ω_j for which $\arg\max_j \left[\frac{1}{L}\sum_{l=1}^{L} p_l(\omega_j|\mathbf{x})\right]$; that is, the class predicted by the majority of the classifiers. We refer to this as a consensus of certainty, and is a way of exploiting the uncorrelated variance in each of the individual classifiers.

However, in some cases it is desirable to know when there is no consensus, or more specifically when the ensemble cannot return a confident classification. Here we are not interested in knowing whether weak learners agree or disagree about the class of $\mathbf{x}$, but rather about the degree of confidence the weak learners have in assigning $\mathbf{x}$ to one of ω_j, $j \in \{1, \cdots, c\}$. The problem may be restated to ask whether $\mathbf{x}$ should belong to an "ambiguous" class or not, where ambiguousness refers to the difficulty (or lack of confidence) in classifying a sample.

4.2.2 Active Learning for Cost-Effective Training

Active Learning (AL) is a method of intelligently training a classifier, mitigating several drawbacks of the more standard Random Learning (RL), where samples are randomly selected for labeling [55]. RL assumes that large amounts of labeled data are already available, but for biomedical domains, manual labeling is costly and time-consuming. For example, digital images of pathology slides can be several gigabytes in size. To build a classifier to detect disease in these images, an expert pathologist needs to provide

precise annotation of disease extent in the image. This results in a large training cost if RL is employed. In contrast, AL selects samples from an unlabeled pool for annotation based on the ambiguity of a sample: samples that are difficult to classify are not currently well-represented within the training set, so by targeting these samples, fewer training samples are needed to achieve high accuracy. Thus by finding only the most difficult to classify samples, we identify the most critical for labeling and inclusion in the training set.

4.2.3 Current Active Learning Approaches

There are several AL methods for selecting training samples [55, 69, 51], each relying upon a single measurement of ambiguity. The Query-By-Committee (QBC) method by Seung, et al. [51] trains a group of L weak learners, each of which votes on the class of sample $\mathbf{x}$. In the two-class case, if the sample receives approximately $\frac{L}{2}$ votes for both classes, then $\mathbf{x}$ is considered ambiguous (difficult to classify). Li, et al. [69] utilized a support-vector machine approach, whereby samples appearing close to a decision hyperplane in high-dimensional space are considered ambiguous. There is no guarantee that each of these methods will identify the same samples as "difficult to classify," since samples that are close to a decision hyperplane may still be unanimously identified as a single class by a QBC algorithm. Thus, the set of ambiguous samples may depend heavily on the AL method.

4.2.4 Novel Contributions of This Paper

In this paper, we present the concept of a consensus of ambiguity (CoA) whereby several measures of ambiguity are combined to identify the most difficult to classify samples from an unlabeled pool. This framework extends beyond the traditional AL methods by identifying ambiguousness explicitly rather than as a function of classification error. We define a consensus ratio that measures the degree of overlap between multiple algorithms for finding ambiguity, and we find that using multiple algorithms ensures that the overlap between methods decreases; the use of multiple algorithms ensures that only the most difficult to classify samples are detected by the algorithm.

We evaluate the efficacy of the algorithm by using the CoA-based AL method to train a probabilistic boosting tree (PBT) classifier on three separate medical image datasets. We use the performance of the PBT, measured in terms of accuracy and area under the receiver operating characteristic curve (AUC), to ensure that the training set created by CoA-based AL can yield higher performance compared to a randomly-selected training set of equal size. The three datasets considered in this work are: (1) Digitized prostate histopathology (100 images) are broken up into 12,000 image regions, each of which is classified as cancer / non-cancer using texture features. (2) 18 prostate dynamic contrast-enhanced MRI (DCE-MRI) images (256x256 pixels) are quantified using textural and functional intensity features to find cancer in a pixel-wise fashion. (3) 9,000 regions of interest (ROIs) are extracted from two large breast histopathology patient studies, with each ROI corresponding to either high or low Bloom-Richardson cancer grades. ROIs are quantified by graph-based nuclear architectural features. Each of these datasets represents different modalities, tissues, and features, but all are time-consuming and expensive to annotate; thus, we expect that AL training algorithms can reduce the expense required to obtain reliable training sets versus a random learning scheme.

4.3 Theory of CoA

4.3.1 Active Learning Strategy Overview

We denote by X a set of data containing samples $\mathbf{x} \in X$. Each sample is associated with a class label $y \in \{\omega_1, \omega_2, \cdots, \omega_c\}$. A supervised classifier is denoted $\mathcal{C}(\mathbf{x}) \in \{\omega_1, \omega_2, \cdots, \omega_c\}$. The classifier returns a hypothesis for a sample and is trained on a training set S^{tr} and tested on an independent testing set. The goal of the AL algorithm is to build S^{tr} from a set of unlabeled samples in X. To do this, a training function $\Phi(\mathbf{x})$ returns a measure of ambiguity for $\mathbf{x}$.

Definition 2. *A sample $\mathbf{x} \in X$ is considered ambiguous if $a < \Phi(\mathbf{x}) < b$, where a, b are lower and upper thresholds for Φ, respectively.*

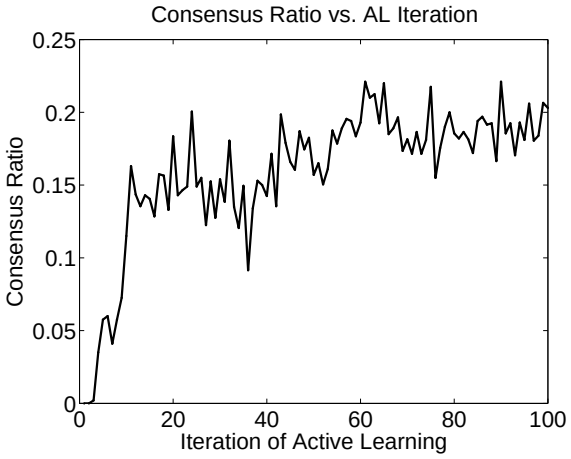


Figure 4.1: Plot of the consensus ratio $\mathcal{R}$ as a function of t, for $t \in \{1, 2, \cdots, 100\}$. After $t = 50$, the consensus ratio plateaus at approximately 0.2. This indicates that there is relatively little consensus between three AL methods: Φ_1 (QBC), Φ_2 (BAY), and Φ_3 (SVD).

4.3.2 Consensus of Ambiguity: Definition and Properties

The CoA approach employs multiple algorithms, $\Phi_1, \Phi_2, \cdots, \Phi_M$, each of which returns a corresponding set of ambiguous samples $S_1^{\mathrm{E}}, S_2^{\mathrm{E}}, \cdots, S_M^{\mathrm{E}}$.

Definition 3. *Given nonempty sets of ambiguous samples, S_i^E, $i \in \{1, \cdots, M\}$, the consensus ratio is defined as $\mathcal{R} = \frac{U}{V}$, where $U = |\bigcap_{i=1}^{M} S_i^E|$ and $V = |\bigcup_{i=1}^{M} S_i^E|$.*

Proposition 3. *Given nonempty sets of ambiguous samples, S_i^E, where $i \in \{1, \cdots, M\}$, $\mathcal{R} = 1$ indicates perfect consensus and $\mathcal{R} = 0$ indicates no consensus across Φ_i.*

Proof. In the case of absolutely no consensus (i.e. no samples are considered ambiguous by all M algorithms), then $\bigcap_{i=1}^{M} S_i^{\mathrm{E}} = \emptyset$, so $\mathcal{R} = 0$. Conversely, when Φ_i, $i \in \{1, \cdots, M\}$ are in perfect agreement (every algorithm identifies exactly the same samples as ambiguous), then $S_1^{\mathrm{E}} = \cdots = S_M^{\mathrm{E}}$, so $\bigcap_{i=1}^{M} S_i^{\mathrm{E}} = \bigcup_{i=1}^{M} S_i^{\mathrm{E}}$ and $\mathcal{R} = 1$. □

Property 1. *When $\mathcal{R} \approx 0$, there is low consensus and high variance among Φ_i, $i \in \{1, \cdots, M\}$.*

This indicates that any agreement among the algorithms will be highly informative and suggesting a benefit to using a consensus approach. Figure 4.1 shows a graph of $\mathcal{R}$ as a function of t, which identifies the iterations of the AL algorithm. Beginning with $t = 0$, the AL algorithm grows a training set by selecting and labeling ambiguous

samples and adding them to the training set. The process iterates for $t \in \{1, \cdots, 100\}$ times in this experiment. Three different AL algorithms were used: QBC, BAY, and SVD (Section 4.4.2). After 50 iterations, $\mathcal{R}$ levels off at approximately 0.2, indicating that there is little consensus among the methods. Thus, a consensus algorithm is likely to be informative.

Definition 4. *A sample* $\mathbf{x} \in X$ *will be considered strongly ambiguous if* $\mathbf{x} \in \widehat{S}^E = \bigcap_{i=1}^{M} S_i^E$*; that is, if the sample is designated as ambiguous by all* Φ_i *for* $i \in \{1, \cdots, M\}$.

Definition 3 is a version of strong ambiguity wherein all M algorithms must select the sample. It is possible that, on any particular AL iteration, no samples will satisfy this criteria. Definition 3 can easily be modified to include samples selected by a majority of algorithms, or any sample identified by more than one algorithm, and so on.

Proposition 4. *As the number of algorithms* Φ_i*,* $i \in \{1, \cdots, M\}$*, being combined increases, the consensus ratio* $\mathcal{R}$ *will monotonically decrease.*

Proof. An added algorithm, denoted Φ_{M+1}, identifies a set of samples denoted S_{M+1}^{E}. If S_{M+1}^{E} is a subset of the current set of ambiguous samples, $\bigcup_{i=1}^{M} S_i^{\mathrm{E}}$, then the denominator of $\mathcal{R}$ does not change since the union will not increase in size. The denominator of $\mathcal{R}$ will decrease, since any elements in $\bigcap_{i=1}^{M} S_i^{\mathrm{E}}$ that are not found in S_{M+1}^{E} will be removed in the new intersection, $\bigcap_{i=1}^{M+1} S_i^{\mathrm{E}}$. Thus $\mathcal{R}$ will decrease in value.

However, if S_{M+1}^{E} contains unique samples not in the current ambiguous sample set, the union will increase in size; that is, $|\bigcup_{i=1}^{M} S_i^{\mathrm{E}}| < |\bigcup_{i=1}^{M+1} S_i^{\mathrm{E}}|$. Thus the denominator of $\mathcal{R}$ will increase. The numerator of $\mathcal{R}$ will not change, since any samples in S_{M+1}^{E} that are not in $\bigcap_{i=1}^{M} S_i^{\mathrm{E}}$ will be removed in the new intersection, $\bigcap_{i=1}^{M+1} S_i^{\mathrm{E}}$. In this case, $\mathcal{R}$ will decrease. □

Property 2. *Adding additional algorithms to the ensemble, will decrease or maintain* $\mathcal{R}$.

By Property 1, ensembles with a low consensus ratio $\mathcal{R}$ ensure that only samples with a very high degree of ambiguity will be identified. Thus increasing M will ensure

that only extremely ambiguous samples are included in $\widehat{S}^{\mathrm{E}}$. However, if $S^{\mathrm{E}}_{M+1} \cap \widehat{S}^{\mathrm{E}} = \emptyset$, then no samples will be considered strongly ambiguous.

4.4 Experimental Setup

4.4.1 Overview of Datasets

Experiment 1 - Prostate cancer on digitized histopathology

Over a million annual prostate biopsies are performed in the US, each of which must be analyzed manually under a microscope [70]. A quantitative system capable of automatically detecting disease can greatly increase the speed and accuracy with which patients are diagnosed for cancer. Digitized glass slides can be over 2 GB in size (several million pixels), with benign and cancer regions appearing close to one another, and so annotation of these samples is difficult. The objective of this experiment is to apply CoA-based AL to build a classifier able to distinguish between cancerous and non-cancerous patches of biopsy tissue.

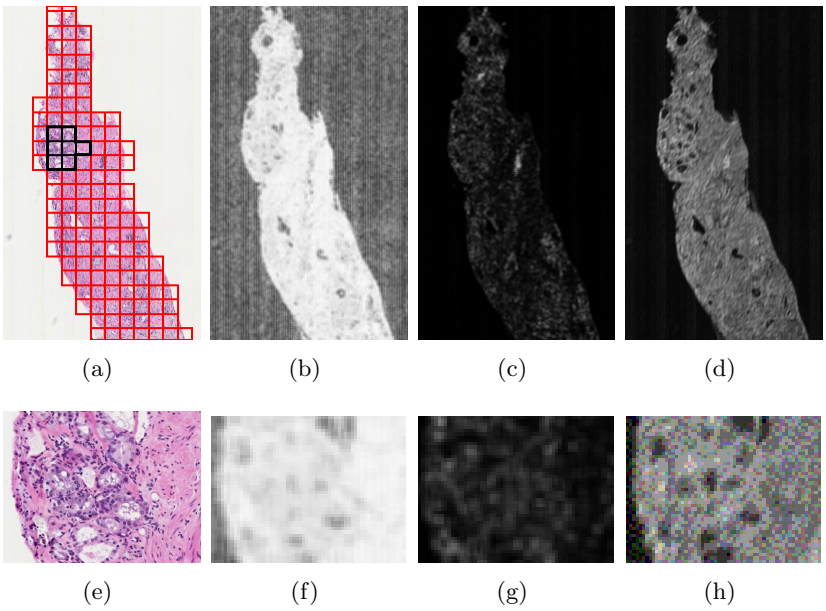


Figure 4.2: Image data from Experiment 1. The original image (a) has a red 30-pixel square grid superimposed, with cancer labeled in black. Texture images are extracted corresponding to first-order greylevel statistics (b), second-order Haralick co-occurrence features (c), and Gabor steerable filter features (d). Shown in the second row (e)-(h) are magnified regions of the cancer region in each image.

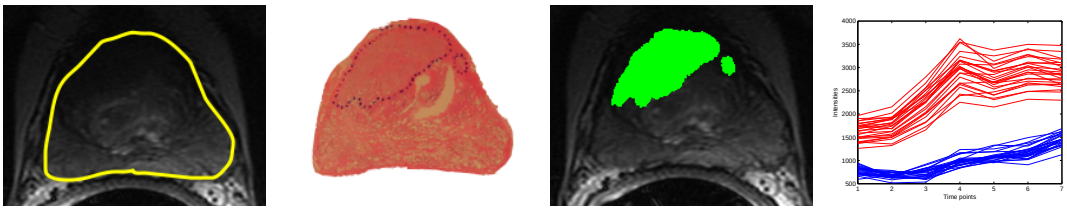
Biopsy samples are stained with Hematoxylin and Eosin (H & E) to visualize cell cytoplasm and nuclei and digitized using a whole-slide digital scanner. For each image, a 30x30 pixel grid is superimposed on the tissue, generating regions of interest (ROIs) of prostate tissue. In previous work [8], we have identified 14 texture features that can easily distinguish between cancer and non-cancer regions of tissue on a pixel-wise basis. These features include: (1) First-order gray-level statistics quantify simple statistics calculated from pixel values in the images [8]. (2) Second-order Haralick features [29] are based on the co-occurrence of pixel values, and are calculated over each ROI. (3) Gabor filter features, also known as steerable filters, operate at a specific orientation and spatial frequency to yield a filter response from the image. Each of the 14 discriminating features is extracted from the image, and the modal value for each 30-by-30 ROI is used as its feature value. 100 images are used to generate 12,000 ROIs which are classified as cancer or non-cancer tissue.

Experiment 2 - Prostate cancer on DCE-MRI

In addition to biopsy, *in vivo* imaging, particularly magnetic resonance imaging (MRI), can be mined for quantitative diagnostic information [34, 71]. Dynamic Contrast Enhanced (DCE) MRI is a technique whereby a contrast agent is injected into a patient with MR images taken at specific time points. The contrast agent is taken up and removed from different tissues at different rates, indicating the presence of disease at a pixel-wise level. A classification system for this modality could be used for automated *in vivo* screening for cancer and treatment, but labeled samples are difficult to obtain since cancer cannot be annotated directly on the MRI. Histopathology is used to find cancer ground truth, which is mapped onto the MR images.

We apply CoA-based AL to a dataset of 6 patients with confirmed prostate cancer on needle biopsies. Prior to radical prostatectomy, MR imaging was performed using an endorectal coil in the axial plane and included T2-w and DCE protocols. Prostatectomy specimens were later sectioned and stained with H & E. An expert pathologist annotated the spatial extent of prostate cancer on the whole-mount prostatectomy sections,

and identified 18 corresponding histopathology and MRI sections. A multimodal registration scheme, COLLection of Image-derived Non-linear Attributes for Registration Using Splines (COLLINARUS) [10], was used to register histology sections onto the corresponding MRI data, thus mapping the cancer ground truth onto the MR images. Structural information from T2-w MRI and functional intensity information from DCE MRI are combined to distinguish between cancer and non-cancer pixels.



(a) (b) (c) (d)

Figure 4.3: Examples of data from Experiment 2. Shown are (a) T2-w MRI image with the prostate boundary in yellow, (b) the corresponding histopathology slice with cancer mapped in blue, and (c) the cancer extent mapped onto the T2-w MRI after registration via COLLINARUS [10]. Also shown are (d) intensity vs. time curves for dynamic contrast; blue curves represent pixel locations in benign tissues, while red curves are inside cancer ground truth ((c)).

Experiment 3 - Breast cancer on digitized histopathology

Breast cancer is the second-leading cause of cancer death in women in the United States [70]. Mammogram screening followed by a biopsy is the current standard for definitive diagnosis. Similar to the motivation in Experiment 1, an automated image analysis system can assist pathologists in detecting and diagnosing breast cancer.

Images of H & E stained breast biopsies are classified between low and high Bloom-Richardson grades of breast cancer. Two patient studies were used to generate 9,000 ROIs of homogeneous tissue measuring 500x500 pixels each. We calculate features based on the architecture of the cell nuclei, in accordance with the major indicators of breast cancer grade. Color deconvolution is used to transform the RGB color space of the image into an alternate three-color space to separate out the hematoxylin, eosin, and white background of the image [19]. Using the deconvoluted image, the centroids of cell nuclei are detected, which are used to construct a series of graphs based on the Voronoi tesselation, Delaunay triangulation, and a minimum spanning tree. From each

of these, a set of quantitative features is extracted to characterize the cell architecture [19]. Each ROI is classified as high or low Bloom-Richardson grades of cancer, where ground truth is determined by a pathologist.

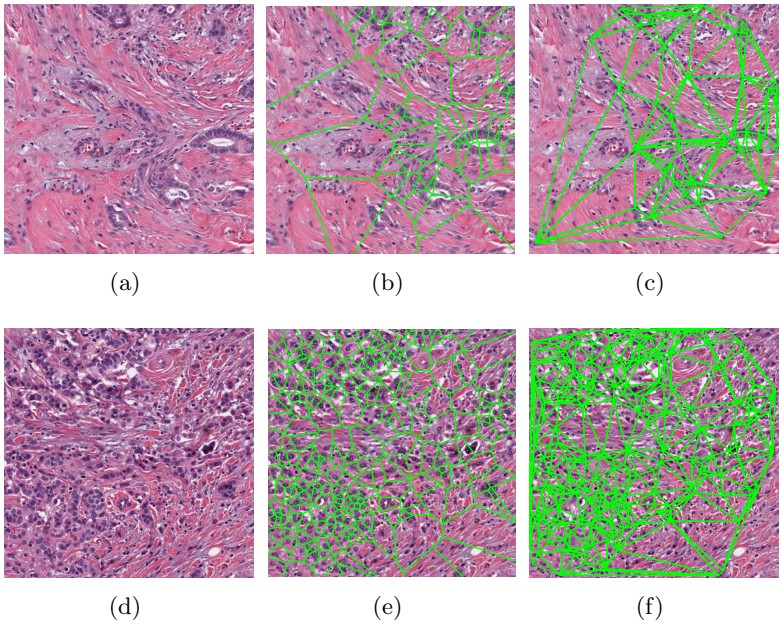


Figure 4.4: Examples of image data from Experiment 3, where we distinguish low-grade breast cancer tissue ((a)-(c)) from high-grade tissue ((d)-(f)). Nuclei are detected from breast biopsy tissue (a), (d) and used to generate graphs such as the Voronoi tesselation (b), (e) and Delaunay triangulation (c), (f). Features from these graphs are used to quantify each image patch.

4.4.2 Comparison of AL Methods

Query-By-Committee (QBC)

QBC [51] involves a group of L weak classifiers that produce votes for the class of an unlabeled sample $\mathbf{x}$. Samples with approximately $\frac{L}{2}$ votes are considered difficult to classify. The output of $\Phi_1(\mathbf{x})$ is the number of votes for the target class, and a, b represent the minimum and maximum votes, respectively. A total of $L = 10$ Random Forests were generated using C4.5 decision trees [36, 45] with threshold values of $a = 4$ and $b = 6$.

Bayes Likelihood (BAY)

Bayes Theorem [41] models the likelihood of observing a class based on the feature values of sample $\mathbf{x}$. A probability density function is created for each of K features, where $p_k(\omega_j|\mathbf{x})$ denotes the likelihood that $\mathbf{x}$ belongs to class ω_j given feature k. Samples for which $p_k(\omega_j|\mathbf{x}) \approx 0.5$ are considered ambiguous. The output of $\Phi_2(\mathbf{x})$ is $\frac{1}{K}\sum_{k=1}^{K} p_k(\omega_1|\mathbf{x})$ where ω_1 is the target (cancer) class. Threshold parameters were set to $a = 0.4$ and $b = 0.6$.

Support Vector Distance (SVD)

Support Vector Machines (SVMs) [37] create a high-dimensional projection of feature data, in which a decision hyperplane is created via training. Samples are classified by finding the position relative to the hyperplane. The output of $\Phi_3(\mathbf{x})$ is the signed distance between $\mathbf{x}$ and the hyperplane, where the sign indicates class membership. Parameters a and b define the distances within which a sample is considered ambiguous. We set a and b to $\pm 10\%$ of the maximum distance from the support vector.

4.4.3 Probabilistic Boosting Tree Classification Algorithm

CoA-based AL was used to train a probabilistic boosting tree (PBT) [68]. The PBT combines AdaBoost [9] and decision trees [45], iteratively generating a tree where each node is boosted with L weak classifiers and whose output is a likelihood for the class of sample $\mathbf{x}$. The PBT algorithm was chosen as a classifier that is different from the methods used in each of the AL algorithms described above. At each iteration of the active learning algorithm, $t \in \{1, 2, \cdots, 100\}$, ambiguous samples found by the CoA ensemble are sampled to obtain equal numbers of samples from both classes [7], which are used to train the PBT. For our experiments, each iteration added two samples (one from each class) to the growing training set. Evaluation on an independent testing set is done via area under the receiver operating characteristic curve (AUC) and accuracy.

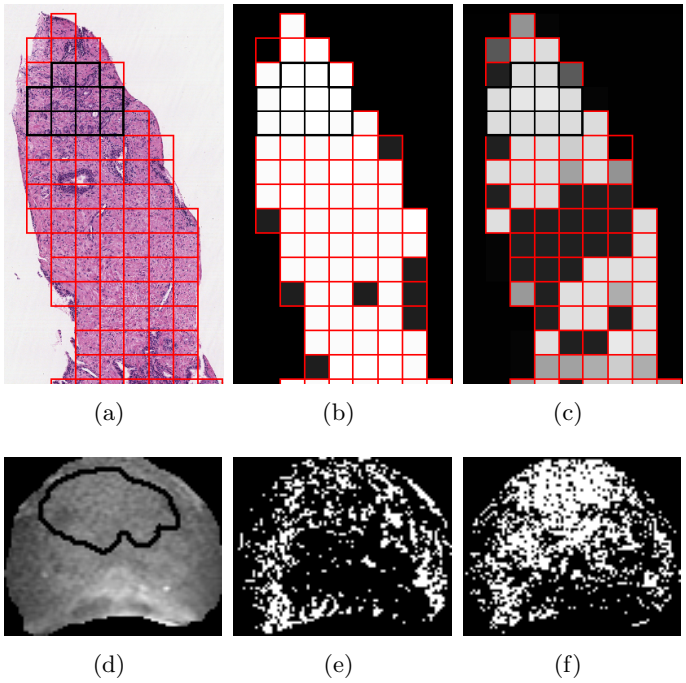


Figure 4.5: Examples of images taken from the prostate histopathology (a) and DCE-MRI (d) datasets, with cancer regions indicated by black contours. Also shown are the corresponding classification results of the PBT, when using training sets built via RL ((b), (e)) and CoA-based AL ((c), (f)). Images were obtained at AL iteration $t = 50$.

4.5 Results and Discussion

Shown in Figure 4.5 are examples of two datasets, prostate histopathology (top row) and DCE-MRI (bottom row), used in this study. In the left column (Figures 4.5 (a), (d)) are the original images with the cancerous region delineated in a black contour, while the results of classification with RL training are shown in the middle column (Figures 4.5 (b), (e)) and training with CoA-based AL are shown in the right column (Figures 4.5 (c), (f)). The images were obtained when the AL algorithm had run for $t = 50$ iterations.

For histopathology, brighter regions indicate higher likelihood of cancer. The RL-trained classifier identifies the majority of patches as cancer yielding a high false-positive count, while the CoA-trained classifier is able to discriminate between obviously benign regions and cancerous areas. Note that we are not commenting here on the accuracy of the final classifier, but on the performance of one training method with respect to another. For the DCE images, images were thresholded at a likelihood of 75%. Here,

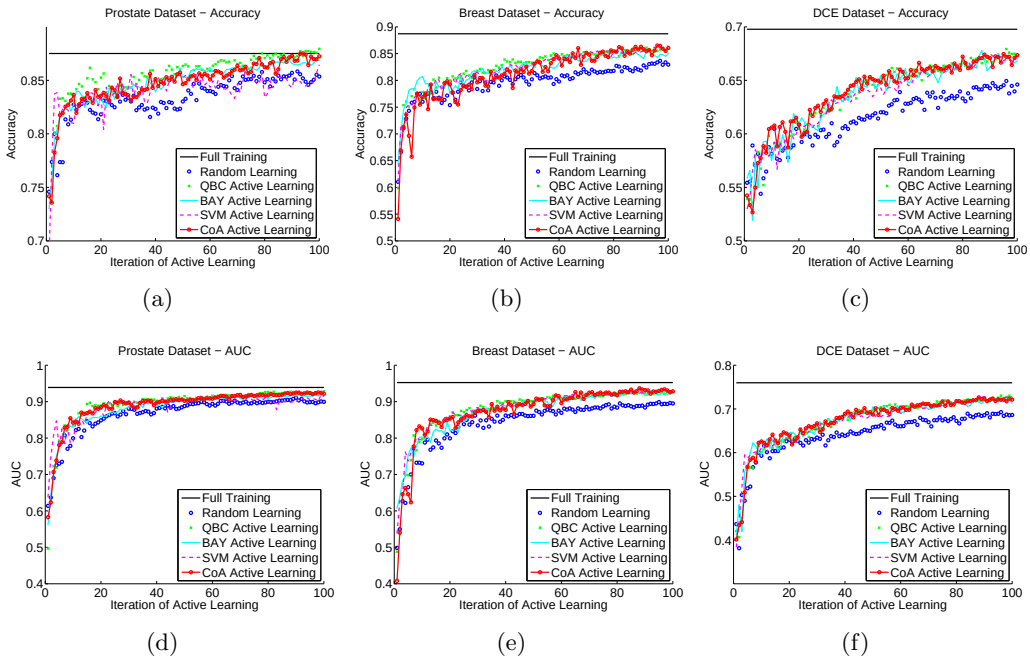


Figure 4.6: Plots of the accuracy and AUC obtained by the PBT using the training derived from CoA Active Learning method (red solid line), which combines three AL schemes (QBC, BAY, and SVD), and Random Learning (blue dotted line). Shown are results for the dataset of 12,000 prostate histopathology ROIs ((a), (d)), 28,000 prostate DCE-MRI pixel samples ((c), (f)), and 9,000 breast histopathology ROIs ((b), (e)).

the RL-trained classifier yields false-negatives with a small set of pixels classified as cancer, while the CoA-trained classifier correctly classifies many pixels near the ground truth. Again, this indicates that – given the limitations on labeling biomedical images – CoA yields better results than random training on a limited number of training samples.

The accuracy and AUC of the PBT are plotted against the AL iteration $t \in \{1, \cdots, 100\}$ in Figure 4.6. Shown are the results for the classifier trained using the CoA algorithm (red solid) as well as random learning (blue dotted) and each of the three AL strategies: QBC (green dot), BAY (cyan solid), and SVM (magenta dash). Each location on the independent axis indicates a training set size (increasing from left to right); we can see that for the majority of training set sizes, all of the AL-trained classifiers yield better accuracy and AUC than random learning. Additionally, AL requires fewer samples to reach that desired performance compared with RL. We note that the individual AL algorithms do not necessarily perform better than the CoA approach in terms of classifier performance, but this is not an unexpected result. The goal of using

the CoA algorithm is to prune down the number of samples deemed "eligible" at each stage; we see that by constraining our search in this way, we have a smaller pool from which to choose labeled samples, while keeping performance the same as an individual algorithm (which has a much wider set of "eligible" samples).

4.6 Concluding Remarks

In this paper, we presented a CoA framework for identifying ambiguousness in an unlabeled pool of data. The CoA approach exploits variance between different ambiguity measurements. A consensus ratio determines the amount of variance between multiple ambiguity methods, and by combining these algorithms, this ratio decreases. This ensures that only the most ambiguous samples are selected from the unlabeled data. Finally, we applied CoA to the problem of Active Learning (AL), where ambiguous samples are selected for training a classifier. For medical image datasets (which are time-consuming and expensive to annotate), the CoA-trained classifier yields higher accuracy and AUC than RL for similar training set sizes.

We observe similar classification performance using CoA versus individual AL training schemes. However, the low consensus ratio indicates that each training algorithm is selecting mostly unique samples. Since our goal is to improve training efficiency, we wish to explore evaluation measures besides classifier performance. For example, it is possible that samples selected by one AL scheme are more difficult to annotate than those selected by another, or have significantly different feature distributions. If so, we may be able to derive an evaluation metric that is divorced from classifier performance that is able to identify the most efficient training algorithm.

Chapter 5

Integrating Manifold Learning with Graph, Textural, and Morphological Features for Automated Grading of Prostate Histology

5.1 Abstract

This paper presents a manifold learning based algorithm which utilizes nearly 600 automatically extracted texture-, graph-, and shape-based features in order to quantitatively grade prostate cancer (CAP) from digitized histology. Currently, pathologists rely on the Gleason grading scheme, a lexicon of visual pathological attributes, in order to identify prostate cancer grades between 1 and 5. The 4 classes of features that we consider seek to capture nuclear and glandular arrangement, morphology, and tissue architecture of the different prostate cancer grades as defined in the Gleason paradigm. We also define a number of graph-based and textural attributes to characterize CAP appearance which fall outside the purview of the Gleason scheme. A manifold learning scheme, spectral clustering, is used to reduce the dimensionality of the feature set in order to visualize inter-class relationships between the intermediate and clinically most difficult to distinguish CAP grades (3, 4). Support vector machine (SVM) and decision tree classifiers then use the results of spectral clustering to automatically distinguish a total of 31 digitized tissue sections as either grade 3, grade 4, or benign epithelium. The SVM achieved an accuracy of 95.8%, while the decision tree obtained an accuracy of 87.5% in distinguishing the 3 tissue classes. Our results further suggest that by combining features explicitly modeled on the Gleason scheme with features that lie outside it we achieve a higher classification accuracy than might have been attainable via Gleason derived features alone.

5.2 Introduction

Prostate cancer is the most commonly diagnosed cancer among males in the U.S., with 200,000 new cases and 27,000 deaths predicted for 2007 (source: *American Cancer Society*). Successful treatment of cancer depends upon reliable methods of detection and tissue analysis. Currently manual examination of prostate biopsy samples under a microscope by an expert pathologist is the gold standard of prostate cancer diagnosis and grading. Regions exhibiting cancer activity are assigned a number to characterize the degree of malignancy found within the tissue sample. This characterization is critical in determining the best course of treatment for a patient. In the U.S., the most common system of numbering or "grading" prostate tissue is the Gleason scale [11], which assigns grades on a scale from 1 (well-differentiated, relatively benign tissue) to 5 (non-differentiated tissue, highly invasive cancer).

The Gleason paradigm, shown in Fig. 5.1, illustrates how cancer grades differ in terms of their tissue appearance. Architecture refers to the spatial arrangement of nuclei and glands within the tissue with respect to their centers of mass, and morphology refers to the shape and size of glands and nuclei. Glands and nuclei both express architectural and morphological changes as cancer progresses from benign to malignant: nuclear proliferation and infiltration increase, glands in the prostate tissue become smaller, circular, and more uniform, and the overall texture qualities of the digital histological images are altered [72]. Examples of tissue regions of Gleason grades 3 and 4 and glands and nuclei from those regions are shown in Fig. 5.2. Recently, studies have identified a number of issues with the Gleason system, including (1) a high degree of observer variability, with tissue under-grading (assigning an incorrectly low grade to a tissue sample) as high as 48% [15], (2) the use of only tissue and glandular architecture features while ignoring global texture and nuclear and glandular contour

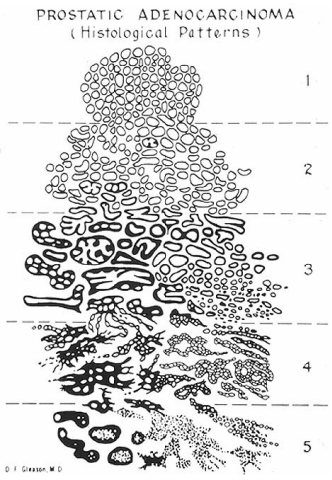
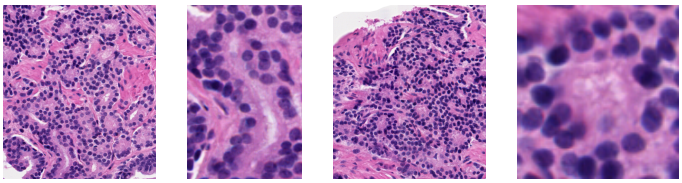


Figure 5.1: Sketch of the Gleason grading system [11] according to tissue patterns.



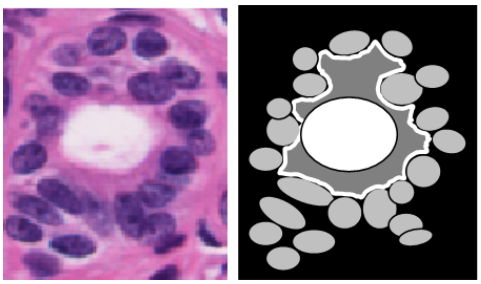
(a) (b) (c) (d)

Figure 5.2: Examples of (a) Gleason grade 3 tissue, (b) a gland from (a) magnified, (c) Gleason grade 4 tissue, and (d) a gland from (c) magnified.

descriptors [72], which may provide additional complementary discriminatory information, and (3) qualitative grouping of CAP into 5 classes, thereby ignoring potential intermediate cancer classes.

Previous attempts at quantitative prostate cancer grading have been undertaken by Roula, et al. [73] using texture and morphological features extracted from multispectral images of prostate biopsy samples. Tissues were classified as benign stroma, hyperplasia, intraepithelial neoplasia, or CAP using Haralick texture and nuclear and glandular morphology features. Tabesh, et al. [23], used color-channel histograms, fractal dimensions, wavelet features, and morphology to distinguish high and low grade CAP. Note that distinguishing between low and high cancer grades is relatively simpler compared to distinguishing between intermediate cancer grades, (3 and 4), which is the more clinically relevant problem [15] (Figs. 5.1 (b) and (d)).

In [74], we presented a computer-aided diagnosis (CAD) system to detect cancerous areas on digitized prostate histology via a hierarchical multiscale framework. In this work, we investigate the use of a manifold learning scheme called spectral clustering with nearly 600 different features corresponding to graph-, morphology-, and texture-based features in order to distinguish intermediate Gleason grades (3 and 4) and benign epithelium. While some of these features (morphology) were modeled on the Gleason scheme, other features (texture and graph-based) were designed outside the purview of the Gleason scheme. The spectral clustering algorithm is employed to non-linearly reduce the dimensionality of the feature space to enable (a) use of a classifier (support vector machines and C4.5 decision trees) to distinguish the different tissue classes and (b) visualization of inter-class relationships.



(a) (b)

Figure 5.3: Example of (a) a gland, and (b) illustration of the lumen and nuclei structures comprising the gland in (a).

In Sections 2 and 3 we describe our gland segmentation and feature extraction methods. The manifold learning algorithm is described in Section 4. Results are presented in Section 5 and concluding remarks in Section 6.

5.3 Level Set Segmentation to Extract Gland Margins

A level set algorithm was used [75] to segment the gland margin (Figure 5.3 (a)) to derive gland boundary features. We denote a tissue region R by a digital image $\mathcal{C}^R = (C, f)$ where C is a 2D grid of image pixels $c \in C$ and f is a function that assigns an intensity to c. A boundary $\mathcal{B}$ evolving in time t and 2D space C is represented by the zero level set $\mathcal{B} = \{(x, y)|\phi(t, x, y) = 0\}$ of a level set function ϕ, where x and y are 2D Cartesian coordinates of $c \in C$. The evolution of ϕ is described by

$$\frac{\partial \phi}{\partial t} + F|\nabla \phi| = 0 \tag{5.1}$$

where function F defines the speed of the evolution, which in turn depends on $f(c)$. The initial contour $\phi_0 = \phi(0, x, y)$ is initialized by the user inside the gland near the lumen area (the white region in Fig. 5.3 (b)) and is allowed to evolve to its final position (white line in Fig. 5.3 (b)). $\mathcal{C}^R$ comprises k glands with centroids at manually labeled pixels $c_g^1, c_g^2, \cdots, c_g^k$. $\mathcal{C}^R$ also comprises m nuclei (grey ellipsoids in Fig. 5.3 (b)) with centroids at manually labeled pixels $c_n^1, c_n^2, \cdots, c_n^m$. Finally, the lumen area (white region in Fig. 5.3 (b)) is segmented via a Bayseian classifier trained on normalized RGB values determined from an independent training set of lumen pixels.

Feature Type	Number of features	Gleason Feature
Nuclear arrangement	25	Nuclear proliferation and infiltration
Graph Features	24	None
Gland Morphology and Architecture	44	Gland differentiation and organization
Texture (Statistical, Haralick, Gabor)	483	None

Figure 5.4: Table describing the groups of features, the number of features in each group, and their relation to the Gleason grading scheme.

5.4 Feature Extraction

Hematoxylin and eosin stained prostate biopsy cores are imaged on a high resolution whole slide digital scanner at 40x magnification and saved on a computer workstation. A total of 31 studies were identified as Gleason grade 3, grade 4, and benign epithelium by an expert pathologist. We considered 4 classes of image features including: (1) nuclear, to quantify the quantity and density of nuclei in the tissue, (2) graph-based, to quantify the spatial arrangement of nuclei, (3) morphological, to quantify the location and shape of glands, and (4) textural, to quantify the pattern and arrangement of structures within the tissue.

5.4.1 Nuclear Features

We compute the following 25 features directly from the spatial location of the centroids of the nuclei in $\mathcal{C}^R$ to characterize nuclear proliferation. (1) The density of the nuclei in $\mathcal{C}^R$ is computed as $\mathcal{D} = \frac{m}{|C|}$, where $|C|$ is the cardinality of C. (2) We denote by $\mathcal{S}_K$ the set of K-nearest neighbors of nuclear centroid c_n^a where $K \in \{3, 5, 7\}$ and $a \in \{1, 2, \cdots, m\}$. Average nuclear distance of c_n^a is given by $d_{c_n^a,K} = \frac{1}{|\mathcal{S}_K|} \sum_{c \in \mathcal{S}_K} ||c_n^a - c||$. The overall average nuclear distance $\mu_{n,K}^d = \frac{1}{m} \sum_a d_{c_n^a,K}$ and standard deviation $\delta_{n,K}^d$ over all $a \in \{1, 2, \cdots, m\}$ is calculated. In addition, a measurement of disorder quantifying the variation of $d_{c_n^a,K}$ for all a is given as $\Psi_{n,K}^d = 1 - (1/(1 + \frac{\mu_{n,K}^d}{\delta_{n,K}^d}))$, giving an additional 9 features for $\mathcal{C}^R$. (3) We denote by $B_{c_n^a,r}$ a ball of pixels with radius r centered on c_n^a. The number of pixels corresponding to nuclear centroids c_n^j, $j \neq a$, $j \in \{1, 2, \cdots, m\}$ in $B_{c_n^a,r}$ are counted and the sum denoted as $Q_{c_n^a,r}$. The mean and

standard deviation of $Q_{c_n^a,r}$ for $a \in \{1, 2, \cdots, m\}$ are denoted by $\mu_{n,r}^Q$ and $\delta_{n,r}^Q$. The measurement of disorder $\Psi_{n,r}^Q$ is also calculated as described above for $\Psi_{n,K}^d$. In this study, we use values of $r \in \{10, 20, \cdots, 50\}$ which were determined empirically.

5.4.2 Graph-based Features to Describe Tissue Architecture

We use different graph algorithms to quantitatively describe the arrangement of nuclei within CAP. From these graphs, we extract an additional 24 features.

Voronoi Diagram

The Voronoi diagram $\mathbf{V}$ partitions $\mathcal{C}^R$ with a series of polygons. A polygon set $P_{c_n^a}$ centered on c_n^a contains any pixel $c \in C$ for which $\mathbf{d}(c, c_n^a) = \min_j\{||c - c_n^j||\}$ where $a, j \in \{1, 2, \cdots, m\}$ and $\mathbf{d}(b, c)$ is the Euclidean distance between any $b, c \in C$. These polygons create a tesselation of $\mathcal{C}^R$ where every pixel is assigned to a polygon and every polygon is associated with a nuclear centroid. Each $P_{c_n^a}$ has e unique edges $E_{b,b+1}^{\mathbf{V}}, E_{b+1,b+2}^{\mathbf{V}}, \cdots, E_{e,b}^{\mathbf{V}}$ between all adjacent vertices with corresponding edge lengths $l_b^{\mathbf{V}}, l_{b+1}^{\mathbf{V}}, \cdots, l_e^{\mathbf{V}}$ and chord lengths $H_1, H_2, \cdots, H_h$ between all nonadjacent vertices. Each $P_{c_n^a}$ has a perimeter $l_E^{\mathbf{V}} = \sum_{i=1}^{e} l_i^{\mathbf{V}}$, total chord length $l_H^{\mathbf{V}} = \sum_{i=1}^{h} H_i$, and total area $A^{\mathbf{V}} = |P_{c_n^a}|$. We compute the average, standard deviation, ratio of minimum to maximum value, and disorder for $A^{\mathcal{V}}$, $l_E^{\mathcal{V}}$, and $l_E^{\mathcal{V}}$ of each $P_{c_n^a}$ in $\mathcal{C}^R$, giving 12 features.

Delaunay Triangulation

The Delaunay graph $\mathbf{D}$ is a graph constructed so that any two unique nuclear centroids c_n^a and c_n^b, where $a, b \in \{1, 2, \cdots, m\}$, are connected by an edge $E_{a,b}^{\mathbf{D}}$ if their associated polygons in $\mathbf{V}$ share a side. The average, standard deviation, minimum to maximum ratio, and disorder of the areas and edge lengths are computed for all triangles in $\mathbf{D}$, giving 8 features.

Minimum Spanning Tree

A spanning tree $\mathbf{S}$ of $\mathbf{D}$ is a subgraph which connects all c_n^a, $a \in \{1, 2, \cdots, m\}$ together. A single $\mathbf{D}$ can have many $\mathbf{S}$. Weights $\omega_{a,b}^{\mathbf{S}}$ are assigned to each edge $E_{a,b}^{\mathbf{S}}$ in each $\mathbf{S}$

based on the length of $E^{\mathbf{S}}_{a,b}$ in $\mathbf{S}$. The sum of all weights $\omega^{\mathbf{S}}_{a,b}$ in each $\mathbf{S}$ is determined to give the weight $\widehat{\omega}^{\mathbf{S}}$ assigned to each $\mathbf{S}$. The minimum spanning tree (MST) denoted by $\mathbf{S}^{\mathcal{T}}$ has a weight $\widehat{\omega}^{\mathbf{S}^{\mathcal{T}}}$ less than or equal to $\widehat{\omega}^{\mathbf{S}}$ for every other spanning tree $\mathbf{S}$. We compute the average, standard deviation, minimum to maximum ratio, and disorder of the edge lengths in $\mathbf{S}^{\mathcal{T}}$ to obtain an additional 4 and a total of 24 graph-based features.

5.4.3 Gland Architecture and Morphology

Under this class of features we consider descriptors to characterize the boundary appearance and arrangement of glands.

Co-Adjacency Features

We denote as $c^1_g, c^2_g, \cdots, c^k_g$ the centroids of k glands within $\mathcal{C}^R$, and construct a co-adjacency matrix W wherein the value of row u, column v, $W(u,v) = ||c^u_g - c^v_g||$, $u, v \in \{1, 2, \cdots, k\}$, and $W \in \mathbb{R}^{k \times k}$. This matrix describes the inter-gland spatial relationships in a manner similar to the co-occurrence matrix proposed by Haralick [29] to describe the spatial relationships between pixel intensity values. We calculate 13 of Haralick's second-order features from W: angular second moment, contrast, correlation, variance, entropy, sum average, sum variance, sum entropy, difference variance, difference entropy, difference moment, and two measurements of correlation [29].

Morphological Features

We denote as $l_{\mathcal{B}}$ the length of the gland boundary $\mathcal{B}$, obtained via level sets as described in Section 5.3. The distance from the centroid of the gland c_g to boundary pixel $c^{\alpha}_{\mathcal{B}}$ is denoted $\mathbf{d}(c_g, c^{\alpha}_{\mathcal{B}})$, where $c^{\alpha}_{\mathcal{B}} \in \mathcal{B}$. We compute the average and maximum of $\mathbf{d}(c_g, c^{\alpha}_{\mathcal{B}})$ over $\alpha \in \{1, 2, \cdots, \beta\}$. We also obtained the fractal dimension of the gland boundary. We picked intermediate points $c^{\gamma}_{\mathcal{B}} \in \mathcal{B}$ where $\gamma \in \{3, 6, 9\}$ on $\mathcal{B}$ and linearly interpolated between these points to obtain length $l^{\gamma}_{\mathcal{B}}$. The fractal dimensions are obtained as $l_{\mathcal{B}}/l^{\gamma}_{\mathcal{B}}$. The following values are calculated for each gland in $\mathcal{C}^R$: gland area $A_{\mathcal{G}}$, lumen area $A_{\mathcal{L}}$, boundary length $l_{\mathcal{B}}$, number of nuclei surrounding the lumen, and number of layers of nuclei encircling the gland. A number of other features are obtained by considering

ratios and combinations of $A_{\mathcal{G}}$, $A_{\mathcal{L}}$, $l_{\mathcal{B}}$, and $\mathbf{d}(c_g, c_{\mathcal{B}}^{\alpha})$ for $\alpha \in \{1, 2, \cdots, \beta\}$, generating 8 values. The average, standard deviation, and measurement of disorder of these 8 features for all k glands is calculated as described in Section 2.1 to yield 24 features for $\mathcal{C}^R$. We also calculate the standard deviation and variance of $\mathbf{d}(c_g, c_{\mathcal{B}}^{\alpha})$ over $\alpha \in \{1, 2, \cdots, \beta\}$, $l_{\mathcal{B}}/l_{\mathcal{B}}^{\gamma}$, and $(l_{\mathcal{B}}^{\gamma})^2/A_{\mathcal{G}}$ for each gland. Finally, for any point on the boundary $c_{\mathcal{B}}^{\alpha} \in \mathcal{B}$ and its adjacent points $c_{\mathcal{B}}^{\alpha-1}$ and $c_{\mathcal{B}}^{\alpha+1}$, the smoothness factor is calculated as $U_{\alpha} = |\mathbf{d}(c_g, c_{\mathcal{B}}^{\alpha}) - \left(\mathbf{d}(c_g, c_{\mathcal{B}}^{\alpha-1}) + \mathbf{d}(c_g, c_{\mathcal{B}}^{\alpha+1})\right)/2|$. The sum of U_{α} for $\alpha \in \{1, 2, \cdots, \beta\}$ is calculated for each gland. The average of these 7 values are computed over k glands in $\mathcal{C}^R$ giving 7 more features for a total of 44 features quantifying gland architecture and morphology.

5.4.4 Texture

First-order Statistics

The average, median, standard deviation, and the range of $f(c)$ is computed for all $c \in C$ for each of the three color channels in the image, yielding 12 first-order statistical features.

Co-occurrence Features

A co-occurrence matrix $\mathcal{Z} \in \mathbb{R}^{\mathcal{M} \times \mathcal{M}}$ is constructed for $\mathcal{C}^R$ where $\mathcal{M} = \max_{c \in C}[f(c)]$ is the maximum pixel value of C. The value in $\mathcal{Z}(f(b), f(c))$ where $b, c \in C$ is given by the number of times intensities $f(b)$ and $f(c)$ occur within a fixed displacement of each other at any orientation. We construct $\mathcal{Z}$ for a displacement of 1. From $\mathcal{Z}$ we extract a total of 39 Haralick features [29] listed in Section 2.3 A.

Gabor Wavelet Features

The modulating function G for the family of 2D Gabor filters [76] is given as:

$$G(x, y, \theta, \kappa) = e^{-\frac{1}{2}((\frac{x'}{\sigma_x})^2 + (\frac{y'}{\sigma_y})^2)} \cos(2\pi\kappa x'), \tag{5.2}$$

where $x' = x\cos(\theta) + y\sin(\theta)$, $y' = y\cos(\theta) + x\sin(\theta)$, κ is the filter scale factor, θ is the filter phase, σ_x and σ_y are the standard deviations along the X, Y axes, and x and y are

the 2D Cartesian coordinates of each image pixel. We convolved the Gabor kernel with the image at 3 pixel neighborhood sizes (3×3, 5×5, 7×7), for $\kappa \in \{0, 1, \cdots, 6\}$ and $\theta \in \{0, \frac{\pi}{8}, \frac{2\pi}{8}, \cdots, \frac{7\pi}{8}\}$. The table in Fig. 5.4 summarizes the full set of 576 extracted features.

5.5 Manifold Learning

Manifold learning (ML) methods reduce the dimensionality of a data set from N dimensions to M dimensions, where $N >> M$, in order to avoid the *curse of dimensionality* for classification and to visualize inter-, and intra-class relationships in a low dimensional embedding space. We used graph embedding [77], a nonlinear technique which seeks to find an embedding of high-dimensional data in a low-dimensional space. Graph embedding constructs a confusion matrix $\mathcal{Y}$ describing the similarity between any two images $\mathcal{C}_p^R$ and $\mathcal{C}_q^R$ with feature vectors $\mathbf{f}_p$ and $\mathbf{f}_q$, which comprise 576 dimensions, and where $p, q \in \{1, 2, \cdots, \mathcal{N}\}$ and $\mathcal{N}$ is the total number of images in the data set.

$$\mathcal{Y}(p, q) = e^{-||\mathbf{f}_p - \mathbf{f}_q||} \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}. \tag{5.3}$$

The embedding vector $\mathcal{X}$ is obtained from the maximization of the function:

$$\mathcal{E}_{\mathcal{Y}}(\mathcal{X}) = 2\eta \frac{\mathcal{X}^{\mathbf{T}}(D - \mathcal{Y})\mathcal{X}}{\mathcal{X}^{\mathbf{T}} D \mathcal{X}}, \tag{5.4}$$

where $D(p, p) = \sum_q \mathcal{Y}(p, q)$ and $\eta = |\mathcal{N}| - 1$. The M-dimensional embedding space is defined by the eigenvectors corresponding to the smallest M eigenvalues of $(D - \mathcal{Y})\mathcal{X} = \lambda D \mathcal{X}$. For image region $\mathcal{C}^R$ defined by feature vector $\mathbf{f}$, the embedding $\mathcal{X}(\mathcal{C}^R)$ contains the coordinates of $\mathcal{C}^R$ in the embedding space and is given as $\mathcal{X}(\mathcal{C}^R) = \big[w_z(\mathcal{C}^R) | z \in \{1, 2, \cdots, M\}\big]$, where $w_z(\mathcal{C}^R)$ are the z eigenvalues associated with $\mathcal{X}(\mathcal{C}^R)$.

5.6 Results

5.6.1 Quantitative Results

Two classifiers (SVMs and decision trees (C4.5)) were tested on 6 Gleason grade 4 (G4) image regions, 18 Gleason grade 3 (G3) regions, and 7 benign epithelial (BE) regions.

Feature Set	G3 vs. G4		G3 vs. BE		G4 vs. BE	
	SVM	*C4.5*	*SVM*	*C4.5*	*SVM*	*C4.5*
Nuclear, Morphology (Gleason derived)	95.8%	91.67%	96.2%	96.0%	92.9%	92.3%
Graph, Texture (Non-Gleason)	87.5%	83.3%	80.8%	84.0%	100%	92.3%
Entire Feature Set	95.8%	87.5%	96.2%	96.0%	100%	92.3%

Figure 5.5: Classification results for Gleason grade 3, grade 4, and benign epithelium tissue regions using SVMs and C4.5 Decision Trees for Gleason (Nuclear, Morphology), non-Gleason (Graph, Texture), and entire feature set.

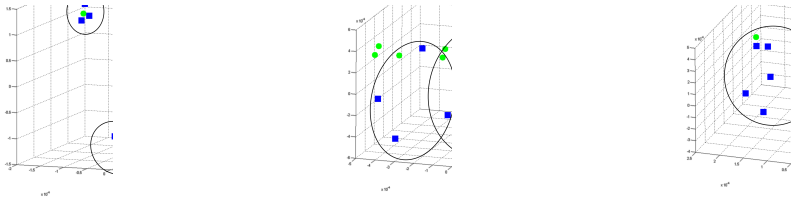
Accuracy results are shown in the table (Figure 5.5). Accuracy for Gleason-based features is higher than for non-Gleason features, but using both sets together perform approximately as well for SVM and C4.5 decision tree classification. We believe that statistically significant differences will become apparent between using Gleason features alone and the entire feature set (the entire feature set performing better) as the size of the image database is increased.

5.6.2 Qualitative Results

Results from manifold learning are shown in Figure 5.6. Feature vectors corresponding to G3 images (green circles) and G4 images (blue squares) are plotted in the 3D embedding space obtained using Gleason-based features (Fig. 5.6 (a)), non-Gleason features (Fig. 5.6 (b)), and both sets combined (Fig. 5.6 (c)). Class clusters, denoted with black contours, tighten and more distinct when using a full feature set (Fig. 5.6 (c)) than either feature set individually.

5.7 Concluding Remarks

In this paper we presented an automated quantitative scheme to distinguish between different grades of prostate cancer on digitized histology. By using a large novel set of 576 features modeled on the Gleason grading paradigm as well as features that fall outside the purview of the Gleason scheme and a novel application of a manifold learning scheme our system achieved an accuracy of over 95% in distinguishing between intermediate cancer grades (3, 4) and benign epithelium. Note that unlike previous approaches that have sought to distinguish between low- and high-grade cancers we



(a) (b) (c)

Figure 5.6: Scatter plots of Gleason grade 3 (green circles) and Gleason grade 4 (blue squares) images, using graph embedding to visualize low-dimensional mappings of (a) nuclear architecture and gland morphology features (non-Gleason), (b) graph and texture-based features (Gleason derived), and (c) all features together.

have focused on the clinically significant problem which is also the reason for inter- and intra-observer grading variability among pathologists, namely distinguishing grades 3 and 4. Our initial results suggest that the integration of features modeled on the Gleason paradigm as well as those that were designed outside its purview, result in higher classification accuracy than attainable by using Gleason derived features alone.

Chapter 6

Manifold Learning for Content-Based Image Retrieval of Prostate Histopathology

6.1 Abstract

We present a content-based image retrieval (CBIR) system for retrieval of digitized images of prostate histopathology. In our study we analyze images from Gleason grade 3, grade 4, and benign epithelium. Nearly 600 image content based features are extracted from digitized images of prostate histology. Manifold learning is used to map the data from a high dimensional non-linear manifold onto a low-dimensional subspace while preserving object adjacency, permitting the use of a linear Euclidean metric for evaluating image similarity. To quantify the efficacy of our CBIR system we analyze a set of 56 digitized prostate histopathology images including 19 benign epithelium, 23 Gleason grade 3, and 14 Gleason grade 4 images. We use feature subsets of 483 textural, 44 morphological, and 49 graph-based features to determine if sub-groups of features were more discriminating than others. The feature space was reduced using 7 different algorithms: principal component analysis, multidimensional scaling, graph embedding, Isomaps, local linear embedding, kernel-based principal component analysis, and laplacian eigenmaps to study the effect of different manifold learning algorithms. The metric was applied to evaluate object similarity in 1 to 10 dimensions in the reduced embedding space. The CBIR system was tested by treating each image as a query image and calculating the similarity of all other images in the reduced feature space using a linear metric. Mean average retrieval precision of 0.573 was obtained for Gleason grade 3, 0.418 for Gleason grade 4, and 0.566 for benign epithelium using morphological features, which performed statistically significantly better than any other subset of features. We found that the highest precision was obtained using principal component

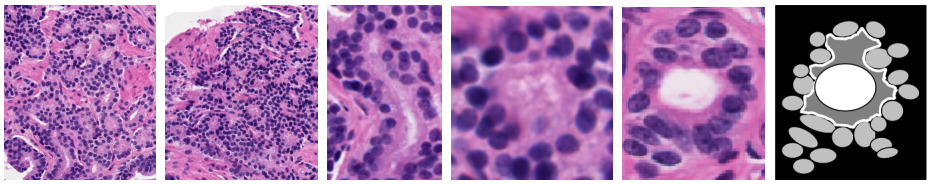
analysis for query images of Gleason grade 3 or grade 4, and laplacian eigenmaps for query images of benign epithelium, reducing the data to less than 3 dimensions for each class.

6.2 Introduction

Prostate cancer is the most commonly diagnosed cancer among males in the U.S., with 200,000 new cases and 27,000 deaths predicted for 2007 (source: *American Cancer Society*). Currently manual examination of prostate biopsy samples under a microscope by an expert pathologist is the gold standard of prostate cancer diagnosis and grading. In the U.S., the most common system of numbering or "grading" prostate tissue (assessing degree of malignancy) is the Gleason scale [11], which assigns grades on a scale from 1 (well-differentiated, relatively benign tissue) to 5 (non-differentiated tissue, highly invasive cancer).

The Gleason paradigm illustrates how cancer grades differ in terms of their architecture (spatial arrangement of nuclei and glands within the tissue with respect to their centers of mass) and morphology (shape and size of glands and nuclei). Glands and nuclei both express architectural and morphological changes as cancer progresses from benign to malignant [11, 72]. An example of tissue regions of Gleason grade 3 tissue is shown in Fig. 5.2 (a), grade 4 tissue in Fig. 5.2 (b), a single grade 3 gland in Fig. 5.2 (c), and a grade 4 gland in Fig. 5.2 (d). A gland from benign epithelial tissue is shown in Fig. 5.2 (e). An illustration of the lumen (white region) and nuclei (grey ellipses) of the gland in Fig. 5.2 (e) is shown in (f). A number of studies have identified issues with the Gleason system, including high degrees of observer variability, with tissue under-grading as high as 48% [15]. Because of the diagnostic importance of Gleason grading, a quantitative system for assisting pathologists in analyzing histopathology will improve patient care by providing an accurate and standardized grading tool.

A great deal of research has focused on creating content-based image retrieval (CBIR) systems to assist physicians in analyzing medical image data [78]. A CBIR system relies on a similarity metric to retrieve images from a database. The metric



(a) (b) (c) (d) (e) (f)

Figure 6.1: Examples of (a) Gleason grade 3 tissue, (b) Gleason grade 4 tissue, (c) a gland from (a) magnified, (d) a gland from (b) magnified, (e) a benign gland, and (f) an illustration of the lumen and nuclei comprising the gland in (e).

used in most systems is a linear distance measure, but because most systems use a large number of features or dimensions, it is common to use manifold learning (ML) methods [78] to map the data into a low-dimensional space. Images that are similar in a high dimensional space will be mapped close together in the transformed space, preserving object similarities. Although many ML methods have been developed over the years, most CBIR systems employ principal component analysis. Zheng, et al. proposed a CBIR system for histopathology in [79] that used color histograms, texture, and Fourier coefficients to describe the content of histological images from various malignancies, using a weighted cosine measure to determine image similarity. However, quantitative evaluation of the system with different feature sets and ML methods was not done.

In [8], we presented a computer-aided diagnosis (CAD) system to detect cancerous areas on digitized prostate histology via a hierarchical multiscale framework. In this work, we present a CBIR system for prostate histopathology. We evaluate the performance of the system by looking at nearly 600 features characterizing texture, morphology, and architecture of histopathological images. We use 7 manifold learning methods to reduce the data to between 1 and 10 different dimensions. The system tested on 56 studies that were identified as Gleason grade 3 (23 studies), grade 4 (14 studies), and benign epithelium (19 studies) by an expert pathologist. The main contributions of this work are:

- A CBIR system for prostate histopathology that employs manifold learning to reduce a selected subset of image features from a high-dimensional, non-linear

manifold to a low-dimensional subspace;

- A novel set of content-based image features that capture characteristics defined in the Gleason scheme (such as morphology) as well as those not analyzed in clinical pathology (such as texture and graph-based features);
- Investigation into the effect of manifold learning algorithm choice and dimensionality on the ability of a CBIR system to retrieve relevant images;

An overview of the CBIR system is described in Section 2. In Section 3 we describe our gland segmentation and feature extraction methods. The manifold learning algorithm is described in Section 4. Results are presented in Section 5 and concluding remarks in Section 6.

6.3 System Overview

An overview of our system is shown in Figure 9.1. Offline, a database of histopathological prostate images is constructed by extracting graph-based, texture, and morphological features from a series of images. These images are then reduced into a low-dimensional space using one of several manifold learning (ML) methods. In this study, we consider principal component analysis (PCA), multidimensional scaling (MDS), graph embedding (GE), Isomaps (ISO), local linear embedding (LLE), kernel-based PCA (k-PCA), and laplacian eigenmaps (LE). Online, a query image is run through the feature extraction algorithm and plotted into the reduced dimensional space using the same ML algorithm that was used in the building of the database. Finally, in the low dimensional space, a linear Euclidean distance metric is used to rank the database images in order of similarity to the query image. The returned images are then output to the user for analysis. A returned image is "relevant" if it is the same class as the query image (Gleason grade 3, grade 4, or benign epithelium), and "irrelevant" otherwise. By comparing MAP values, we can determine the following: (1) which dimensionality reduction algorithm yields the best retrieval precision, (2) the optimal number of dimensions for each ML method, and (3) which feature classes perform best in describing the database and query images in the CBIR system. A Student's t-test

was performed to determine if the difference in performance when using different classes of features is statistically significant.

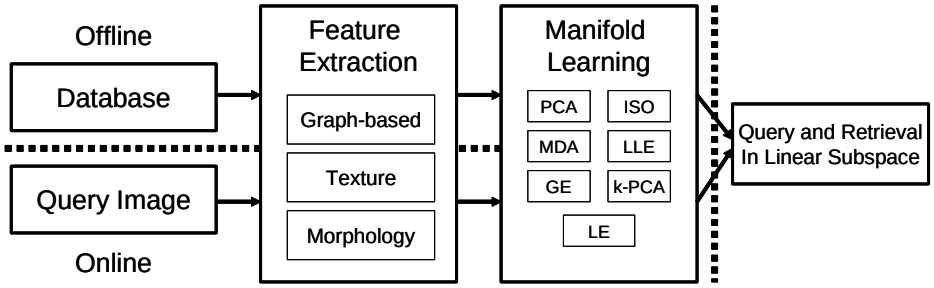


Figure 6.2: Overview and organization of our CBIR system for automated retrieval of prostate histopathology images.

6.4 Feature Extraction

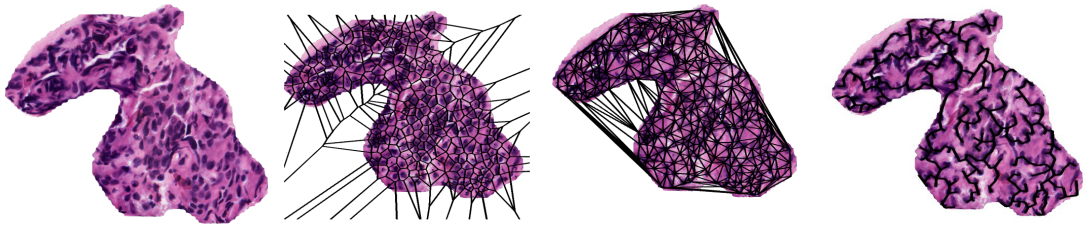
Hematoxylin and eosin stained prostate biopsy cores are imaged on a high resolution whole slide digital scanner at 40x magnification and saved on a computer workstation. We denote a tissue region R by a digital image $\mathcal{C}^R = (C, f)$ where C is a 2D grid of image pixels $c \in C$ and f is a function that assigns an intensity to c. $\mathcal{C}^R$ comprises k glands with centroids at manually labeled pixels $c_g^1, c_g^2, \cdots, c_g^k$. $\mathcal{C}^R$ also comprises m nuclei (grey ellipsoids in Fig. 6.1 (f)) with centroids at manually labeled pixels $c_n^1, c_n^2, \cdots, c_n^m$.

6.4.1 Nuclear Features

We compute the following 25 features directly from the spatial location of the centroids of the nuclei in $\mathcal{C}^R$ to characterize nuclear proliferation. (1) The density of the nuclei in $\mathcal{C}^R$ is computed as $\mathcal{D} = \frac{m}{|C|}$, where $|C|$ is the cardinality of C. (2) We denote by $\mathcal{S}_K$ the set of K-nearest neighbors of nuclear centroid c_n^a where $K \in \{3, 5, 7\}$ and $a \in \{1, 2, \cdots, m\}$. Average nuclear distance of c_n^a is given by $d_{c_n^a,K} = \frac{1}{|\mathcal{S}_K|} \sum_{c \in \mathcal{S}_K} ||c_n^a - c||$. The overall average nuclear distance $\mu_{n,K}^d = \frac{1}{m} \sum_a d_{c_n^a,K}$ and standard deviation $\delta_{n,K}^d$ over all $a \in \{1, 2, \cdots, m\}$ is calculated. In addition, a measurement of disorder quantifying the variation of $d_{c_n^a,K}$ for all a is given as $\Psi_{n,K}^d = 1 - (1/(1 + \frac{\mu_{n,K}^d}{\delta_{n,K}^d}))$, giving an additional 9 features for $\mathcal{C}^R$. (3) We denote by $B_{c_n^a,r}$ a ball of pixels with radius

r centered on c_n^a. The number of pixels corresponding to nuclear centroids c_n^j, $j \neq a$, $j \in \{1, 2, \cdots, m\}$ in $B_{c_n^a, r}$ are counted and the sum denoted as $Q_{c_n^a, r}$. The mean and standard deviation of $Q_{c_n^a, r}$ for $a \in \{1, 2, \cdots, m\}$ are denoted by $\mu_{n,r}^Q$ and $\delta_{n,r}^Q$. The measurement of disorder $\Psi_{n,r}^Q$ is also calculated as described above for $\Psi_{n,K}^d$. In this study, we use values of $r \in \{10, 20, \cdots, 50\}$ which were determined empirically.

6.4.2 Graph-based Features to Describe Tissue Architecture



(a) (b) (c) (d)

Figure 6.3: Examples of graphs superimposed on a patch of Gleason grade 4 tissue (a). Shown are (b) the Voronoi Diagram, (c) the Delaunay Triangulation, and (d) the Minimum Spanning Tree.

Voronoi Diagram

The Voronoi diagram $\mathbf{V}$ partitions $\mathcal{C}^R$ with a series of polygons. Polygon $P_{c_n^a}$ is constructed around c_n^a, creating a tesselation of $\mathcal{C}^R$. Every pixel is assigned to a polygon and every polygon is associated with a nuclear centroid. Each $P_{c_n^a}$ has e unique edges $E_{b,b+1}^{\mathbf{V}}, E_{b+1,b+2}^{\mathbf{V}}, \cdots, E_{e,b}^{\mathbf{V}}$ between all adjacent vertices with corresponding edge lengths $l_b^{\mathbf{V}}, l_{b+1}^{\mathbf{V}}, \cdots, l_e^{\mathbf{V}}$ and chord lengths $H_1, H_2, \cdots, H_h$ between all nonadjacent vertices. Each $P_{c_n^a}$ has a perimeter $l_E^{\mathbf{V}} = \sum_{i=1}^{e} l_i^{\mathbf{V}}$, total chord length $l_H^{\mathbf{V}} = \sum_{i=1}^{h} H_i$, and total area $A^{\mathbf{V}} = |P_{c_n^a}|$. We compute the average, standard deviation, ratio of minimum to maximum value, and disorder for $A^{\mathbf{V}}$, $l_E^{\mathbf{V}}$, and $l_E^{\mathbf{V}}$ of each $P_{c_n^a}$ in $\mathcal{C}^R$, giving 12 features.

Delaunay Triangulation

The Delaunay graph $\mathbf{D}$ is a graph constructed so that any two unique nuclear centroids c_n^a and c_n^b, where $a, b \in \{1, 2, \cdots, m\}$, are connected by an edge $E_{a,b}^{\mathbf{D}}$ if their associated polygons in $\mathbf{V}$ share a side. The average, standard deviation, minimum to maximum

ratio, and disorder of the areas and edge lengths are computed for all triangles in $\mathbf{D}$, giving 8 features.

Minimum Spanning Tree

A spanning tree $\mathbf{S}$ of $\mathbf{D}$ is a subgraph which connects all c_n^a, $a \in \{1, 2, \cdots, m\}$ together. A single $\mathbf{D}$ can have many $\mathbf{S}$. The minimum spanning tree (MST) denoted by $\mathbf{S}^{\mathcal{T}}$ has a total length less than or equal to the total length of every other spanning tree. We compute the average, standard deviation, minimum to maximum ratio, and disorder of the edge lengths in $\mathbf{S}^{\mathcal{T}}$ to obtain an additional 4 and a total of 24 graph-based features.

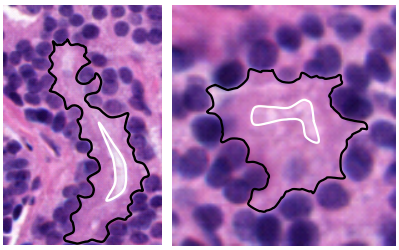
6.4.3 Gland Architecture and Morphology

Co-Adjacency Features

We denote as $c_g^1, c_g^2, \cdots, c_g^k$ the centroids of k glands within $\mathcal{C}^R$, and construct a co-adjacency matrix W wherein the value of row u, column v, $W(u, v) = ||c_g^u - c_g^v||$, $u, v \in \{1, 2, \cdots, k\}$, and $W \in \mathbb{R}^{k \times k}$. This matrix describes the inter-gland spatial relationships in a manner similar to the co-occurrence matrix proposed by Haralick [29] to describe the spatial relationships between pixel intensity values. We calculate 13 of Haralick's second-order features from W: angular second moment, contrast, correlation, variance, entropy, sum average, sum variance, sum entropy, difference variance, difference entropy, difference moment, and two measurements of correlation [29].

Morphological Features

The lumen area is surrounded by a boundary $\mathcal{B}$ obtained via a level-set algorithm [75], where the initial contour is initialized by the user inside the gland near the lumen area (the white region in Fig. 6.1 (f)) and is allowed to evolve to its final position (white line in Fig. 6.1 (f)). We denote as $l_{\mathcal{B}}$ the length of the gland boundary $\mathcal{B}$. The distance from the centroid of the gland c_g to boundary pixel $c_{\mathcal{B}}^{\alpha}$ is denoted $\mathbf{d}(c_g, c_{\mathcal{B}}^{\alpha})$, where $c_{\mathcal{B}}^{\alpha} \in \mathcal{B}$. We compute the average and maximum of $\mathbf{d}(c_g, c_{\mathcal{B}}^{\alpha})$ over $\alpha \in \{1, 2, \cdots, \beta\}$. We also obtained the fractal dimension of the gland boundary. We picked intermediate points



(a) (b)

Figure 6.4: Examples of (a) Gleason grade 3 gland and (b) Gleason grade 4 gland. The lumen boundary is shown in white.

$c_{\mathcal{B}}^{\gamma} \in \mathcal{B}$ where $\gamma \in \{3, 6, 9\}$ on $\mathcal{B}$ and linearly interpolated between these points to obtain length $l_{\mathcal{B}}^{\gamma}$. The fractal dimensions are obtained as $l_{\mathcal{B}}/l_{\mathcal{B}}^{\gamma}$.

The following values are calculated for each gland in $\mathcal{C}^R$: gland area $A_{\mathcal{G}}$, lumen area $A_{\mathcal{L}}$, boundary length $l_{\mathcal{B}}$, number of nuclei surrounding the lumen, and number of layers of nuclei encircling the gland. A number of other features are obtained by considering ratios and combinations of $A_{\mathcal{G}}$, $A_{\mathcal{L}}$, $l_{\mathcal{B}}$, and $\mathbf{d}(c_g, c_{\mathcal{B}}^{\alpha})$ for $\alpha \in \{1, 2, \cdots, \beta\}$, generating 8 values. The average, standard deviation, and measurement of disorder of these 8 features for all k glands is calculated as described in Section 2.1 to yield 24 features for $\mathcal{C}^R$. We also calculate the standard deviation and variance of $\mathbf{d}(c_g, c_{\mathcal{B}}^{\alpha})$ over $\alpha \in \{1, 2, \cdots, \beta\}$, $l_{\mathcal{B}}/l_{\mathcal{B}}^{\gamma}$, and $(l_{\mathcal{B}}^{\gamma})^2/A_{\mathcal{G}}$ for each gland. Finally, for any point on the boundary $c_{\mathcal{B}}^{\alpha} \in \mathcal{B}$ and its adjacent points $c_{\mathcal{B}}^{\alpha-1}$ and $c_{\mathcal{B}}^{\alpha+1}$, the smoothness factor is calculated as $U_{\alpha} = |\mathbf{d}(c_g, c_{\mathcal{B}}^{\alpha}) - \left(\mathbf{d}(c_g, c_{\mathcal{B}}^{\alpha-1}) + \mathbf{d}(c_g, c_{\mathcal{B}}^{\alpha+1})\right)/2|$. The sum of U_{α} for $\alpha \in \{1, 2, \cdots, \beta\}$ is calculated for each gland. The average of these 7 values are computed over k glands in $\mathcal{C}^R$ giving 7 more features for a total of 44 features quantifying gland architecture and morphology.

6.4.4 Texture Descriptors

The average, median, standard deviation, and the range of $f(c)$ is computed for all $c \in C$ for each of the three color channels in the image, yielding 12 first-order statistical features. A co-occurrence matrix $\mathcal{Z} \in \mathbb{R}^{\mathcal{M} \times \mathcal{M}}$ is constructed for $\mathcal{C}^R$ where $\mathcal{M} = \max_{c \in C}[f(c)]$ is the maximum pixel value of C. The value in $\mathcal{Z}(f(b), f(c))$ where $b, c \in C$ is given by the number of times intensities $f(b)$ and $f(c)$ occur within a fixed displacement of each other at any orientation. We construct $\mathcal{Z}$ for a displacement of 1. From $\mathcal{Z}$ we extract a total of 39 Haralick features [29] from each image. Finally, a

family of 2D Gabor filter kernels [27] is created from a modulating function, which is constructed from a Gaussian function modulated by a sinusoid.

6.5 Manifold Learning and Similarity Metric

Manifold learning (ML) methods reduce the dimensionality of a data set from N dimensions to M dimensions, where $M << N$, while preserving the high-dimensional relationships between data points. Since class structure is preserved, ML techniques are employed to avoid the *curse of dimensionality* and to enable the use of a Euclidean similarity metric in a low dimensional embedding space. Many ML techniques have been developed over the years and have been tested on a variety of data sets. Some methods employ a linear algorithm to map the data to a low-dimensional space, while others use a non-linear algorithm, assuming that the data lie on a non-linear manifold in the high-dimensional space. The choice of ML techniques for a particular application is typically arbitrary, since it is difficult to predict which method will produce the best results based on the data. Previously, studies have found that for genome expression data sets, non-linear methods outperform linear methods in mapping out the true class relationships in a low-dimensional space [56]. In this study, we chose to implement the 7 ML methods described in Section 6.3 independently of one another to compare their abilities in distinguishing between different grades of prostate cancer using 3 different subsets of features. For each of these methods, we reduce the dataset to dimensions $M \in \{1, 2, \cdots, 10\}$. Results are generated for each M for comparison. In addition to the ML methods, we analyzed the full data set without reduction to evaluate whether retrieval in the reduced dimensional space is improved compared to the unreduced space.

6.6 Results

Comparing Manifold Learning Methods

By iterating through all of the returned images, we evaluate the system using precision vs. recall (PR) graphs [78], where *precision* is the ratio of the number of relevant images

Feature Type	Query Image	NR	PCA	MDS	GE	ISO	LLE	k-PCA	LE
All Features	Grade 3	0.369	0.392	0.370	0.459	0.37	0.39	0.438	0.493
	Grade 4	0.228	0.267	0.228	0.247	0.228	0.291	0.262	0.261
	Epithelium	0.338	0.471	0.338	0.332	0.339	0.351	0.313	0.334
Texture	Grade 3	0.370	0.444	0.370	0.460	0.370	0.406	0.388	0.468
	Grade 4	0.228	0.268	0.228	0.246	0.228	0.240	0.255	0.273
	Epithelium	0.338	0.425	0.338	0.332	0.339	0.437	0.379	0.356
Graph	Grade 3	0.412	0.445	0.412	0.415	0.412	0.416	0.374	0.454
	Grade 4	0.264	0.270	0.264	0.277	0.264	0.302	0.240	0.317
	Epithelium	0.331	0.405	0.331	0.412	0.331	0.400	0.357	0.373
Morphology	Grade 3	0.567	**0.573**	**0.573**	0.554	0.571	0.557	0.402	0.550
	Grade 4	0.396	**0.417**	0.396	0.385	0.396	0.383	0.229	0.391
	Epithelium	0.462	0.465	0.466	0.518	0.469	0.518	0.484	**0.566**

Table 6.1: Mean average precision values for each queried class. Shown are the highest MAP over $M \in \{1, 2, \cdots, 10\}$. Boldface values are the highest obtained for each class.

Query Image	MDS			GE		
	Texture	**Graph**	**All**	**Texture**	**Graph**	**All**
Gleason Grade 3	6.82E-10	1.63E-09	6.82E-10	5.42E-04	8.13E-07	5.09E-04
Gleason Grade 4	5.68E-05	6.64E-04	5.68E-05	9.00E-04	4.06E-03	9.37E-04
Benign Epithelium	4.01E-03	4.70E-02	4.01E-03	3.44E-04	8.76E-02	3.44E-04

Table 6.2: Results of a two-tailed paired Student's t-test, comparing MAP values for morphology against different subsets of features using two different ML methods. P-values less than 0.05 indicate significantly different results.

retrieved to the total number of retrieved images and *recall* is the ratio of the number of relevant images to the total number of relevant images in the database. A recall of 1.0 is obtained when all images are retrieved from the database, while a precision of 1.0 is obtained if all retrieved images are relevant. The retrieved images are sorted in order of increasing Euclidean distance from the query image, so that the first image returned is most similar to the query. Each image is queried against the remaining images in the database, and we iterate through each of the returned images to generate a PR graph. The PR graphs obtained for all images of the same class are averaged together. We also calculate the *Mean Average Precision* (MAP), an average of the precision for all returned images.

MAP values are shown in Table 6.1. Each row in the table represents the MAP obtained using a particular feature set and class of the query image, and each column shows the ML method used. Because each of the ML methods was used to reduce the data to $M \in \{1, 2, \cdots, 10\}$, the highest MAP values over all M are shown. For each

class, the highest MAP values are shown in boldface. In all three classes, the highest MAP values were obtained when using only morphological features. PCA yielded the highest MAP for Gleason grades 3 and 4, while LE produced the highest MAP for benign epithelium. MDS performed as well as PCA when Gleason grade 3 was the query image, but performance decreased when Gleason grade 4 was the query image. In addition, MAP is highest when the number of dimensions is low (between 1 and 2), suggesting that the majority of the discriminating information is held in only a few dimensions.

Comparing Feature Sets

Table 6.2 shows the results from a two-tailed paired Student's t-test comparing MAP values obtained using morphological features alone to those of the indicated feature subsets. In table 6.2 are shown are the results from two of the ML methods analyzed, MDS and GE. In almost all cases, the values indicate that morphological features result in a statistically significant change in MAP values.

Query Image	**Best Feature Subset**	**Best ML**	**Best M**
Grade 3	Morphology	PCA & MDS	1
Grade 4	Morphology	PCA	2
Benign Epithelium	Morphology	LE	2

Table 6.3: Summary of the best parameters found for each query image class.

Qualitative Results

Results from manifold learning are shown in Figure 6.5. Feature vectors are plotted in the 3-dimensional subspace obtained through (a) MDS and (c) PCA, as applied to morphological features, which performed the best in quantitative analysis. Points in the scatter plot correspond to Gleason grade 3 (green circles), Gleason grade 4 (blue squares), and benign epithelium (red triangles). Class clusters show some separation between the classes when using the reduced feature vectors. Because of their similar appearance, images representing Gleason grades 3 and 4 generally appear very close to

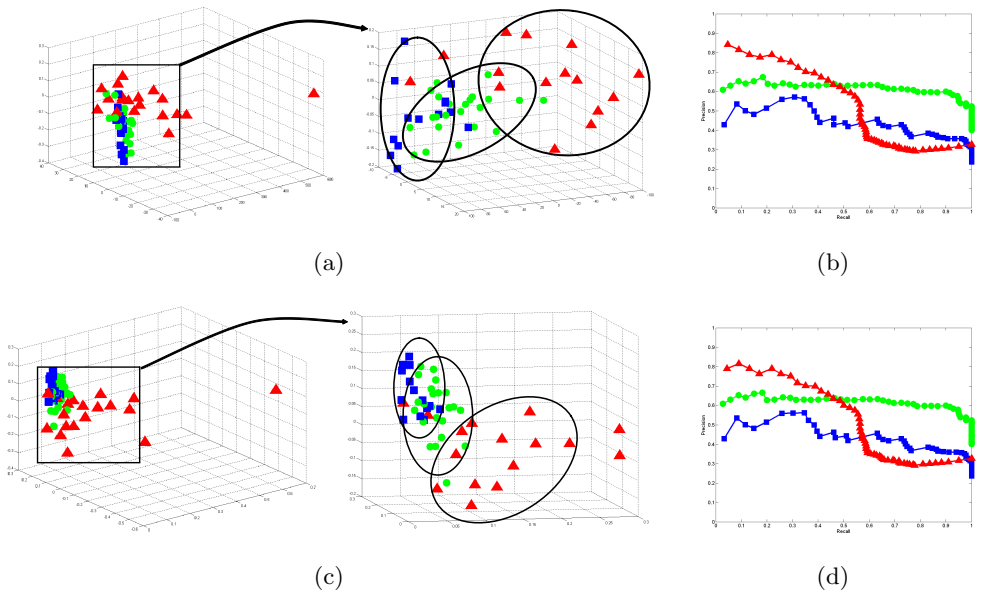


Figure 6.5: Scatter plots obtained through (a) MDS and (c) PCA, with a closeup of the boxed region. The PR curve for all classes obtained using (b) MDS and (d) PCA. Shown are images from Gleason grade 3 (green circles), Gleason grade 4 (blue squares), and benign epithelium (red triangles). Class clusters are manually indicated in black.

one another in the reduced space. The boxed region contains the majority of points and is shown magnified. Precision vs. recall (PR) curves are also shown in Figure 6.5 obtained using (b) MDS and (d) PCA, again using morphological features. For each of the ML methods, the precision of benign epithelium is high for a low number of returned images and decreases as more images are returned, while the Gleason grade 3 images and grade 4 images have consistent precision as the number of retrieved images increases.

6.7 Concluding Remarks

In this paper we presented a CBIR system to retrieve images from a database using a novel set of nearly 600 features. We analyzed 7 different manifold learning methods and 10 lower dimensions. Table 6.3 summarizes the parameters for each class which yielded the best MAP. In this paper, we found:

- Morphological features yield the highest MAP for all classes of query images,

- MAP was highest using a subspace obtained using PCA for Gleason grades 3 and 4 and LE for benign epithelium with dimensions between 1 and 2, and
- Use of ML methods improves retrieval precision over using the unreduced feature space.

Manifold learning normally requires re-computation of the low-dimensional space whenever a sample is added to the dataset (i.e. when a query image must be compared with the database). Law and Jain [80] have proposed a method of incremental nonlinear dimensionality reduction which does not require a re-computation of the embedding, which would allow query images to be quickly compared with the database images in the low dimensional space. Note that unlike previous approaches that have sought to distinguish between low- and high-grade cancers we have focused on the clinically significant problem which is also the reason for inter- and intra-observer grading variability among pathologists, namely distinguishing grades 3 and 4. Our initial results suggest that a CBIR system using features modeled on the Gleason paradigm as well as those that were designed outside its purview can be constructed for the benefit of clinical pathologists. Choice of ML algorithms, number of reduced dimensions, and feature subsets are critical in designing an optimal CBIR system. A comprehensive study of each of these parameters will be possible as more data becomes available for analysis.

Chapter 7

Cascaded Multi-Class Pairwise Approach to Automated Classification of Normal, Cancerous, and Confounder Classes of Prostate Tissue

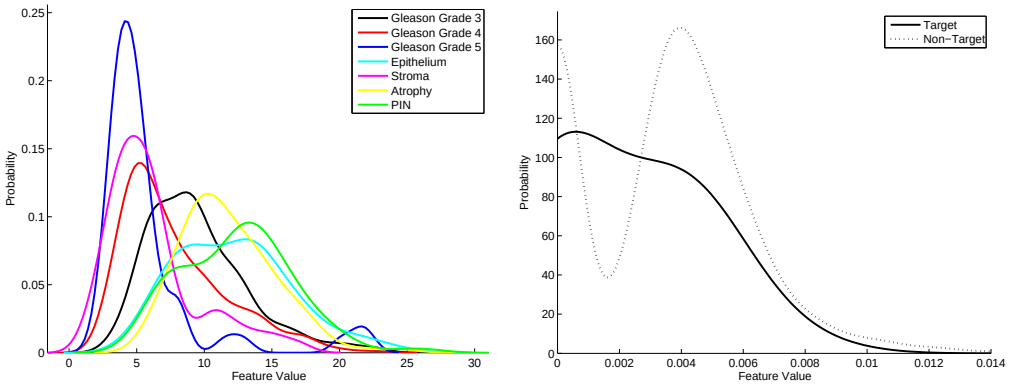
7.1 Abstract

Classification of digitized histopathology traditionally focuses on two-class problems (cancer vs. benign, low vs. high malignancy). However, many tissues consist of several classes, necessitating a multi-class approach. Two common strategies are one-shot classification (OSC), where all classes are identified simultaneously, and one-versus-all (OVA), where a "target" class is distinguished from all "non-target" classes. OSC suffers when several classes of varying similarity must be classified, and OVA relies upon correctly modeling of the very heterogeneous "non-target" class, leading to misclassification on similar samples. In this work, we present a cascaded (CAS) classification approach, where a series of successively granular binary classifications between class groups are performed. The groups are constructed to maximize within-group homogeneity while maximizing between-group heterogeneity, thus avoiding the problem of heterogeneous "non-target" classes in OVA as well as the multiple decision boundaries in OSC. We apply our classification strategy to the problem of classifying 350 regions of tissue taken from digitized prostate biopsy samples (from 214 patients) into six classes: epithelium, stroma, atrophy, prostatic intraepithelial neoplasia (PIN), and Gleason grades 3, 4, and 5. We employ an automated nuclear detection algorithm to extract tissue architectural features as well as texture features from each image, and perform classification using a decision tree algorithm. We find that the CAS strategy achieves an accuracy of 0.64, compared with 0.36 for OVA and 0.34 for OSC, as well as a positive predictive value of 0.64, compared with 0.37 for OVA and 0.34 for OSC.

7.2 Introduction

Prostate cancer is the second most common type of cancer diagnosed among men in the United States, with over 217,000 estimated new cancer diagnoses in 2010 [70]. Over a million prostate biopsies are performed annually in the US, yielding between 12-15 tissue samples per patient that must be analyzed under a microscope for presence of cancer. The current gold standard for cancer diagnosis typically involves manual inspection of biopsy tissue samples under a microscope. The aggressiveness of cancer is determined using the Gleason grading scale, which relies primarily on tissue architecture [11]. There are several drawbacks to this type of manual analysis: (1) Due to the subjective nature of Gleason grading, the system is often applied inconsistently in practice leading to a high degree of inter-observer variability [72]. This is due in part to the existence of tissue classes that act as confounders (non-cancerous conditions that either mimic or are closely associated with cancer growth). However, it may be necessary to correctly identify these confounder tissue types to correctly diagnose and treat a patient [14]. (2) Manual examination of biopsy samples requires a considerable amount of time and effort: each of the 12-15 biopsy samples per patient must be carefully inspected for evidence of cancer. (3) The use of physical tissue samples impedes the consultation of outside experts, and the transport of difficult cases for second-opinion reading is expensive and time-consuming.

Digital pathology (DP) refers to the use of computers to aid in the analysis of tissue samples. Digital slide scanners, which can generate high-resolution images of tissue slides, have become increasingly available and affordable in recent years, and the costs of digital storage and transmission have dropped significantly. In addition, algorithms have been developed to automatically analyze and interpret digital images of tissue samples. Supervised classifiers for digital pathology can be used to reduce variability and improve pathologist agreement [8, 48]. This quantitative analysis can help to mitigate the drawbacks of manual analysis by providing fast, quantitative, and automated analysis of tissue samples. The use of DP solutions in the clinic can facilitate outside consultations and second-readers (telepathology), quantitative database searches



(a) (b)

Figure 7.1: Illustration of various multi-class classification strategies. Shown are probability density functions, where the likelihood of observing a particular class (dependent axis) is plotted against a feature value (independent axis). Shown are two different multi-class strategies: (a) one-shot classification (OSC), where all classes are plotted simultaneously, and (b) one-versus-all (OVA), where a "Target" class is separated all "Non-target" classes.

via content-based image retrieval, and an interactive teaching tool for residents and medical students.

Previous work on supervised classification systems for histopathology have typically focused on two-class problems [23, 81, 82, 83]. However, constructing a two-class classifier (for example, detecting "cancer" versus "non-cancer") results in the inclusion of several heterogeneous tissue types into a single class. A non-cancer class includes epithelium and stroma as well as confounding classes such as atrophy, prostatic intraepithelial neoplasia, and perineural invasion. These non-cancerous classes may contribute to false-positive cancer classification, and these tissue classes may hold important diagnostic and prognostic information themselves [84, 14]. Thus, we may wish to build a multi-class classifier that can accurately distinguish between the variety of different tissue types found in a prostate biopsy.

There are two common approaches to the multi-class problem. The first is to perform one-shot classification (OSC) of several classes at once. These typically involve classifiers that inherently deal with multiple possible class labels, such as decision trees [45]. This approach suffers when dealing with multiple similar classes, since all classes must be distinguished simultaneously using the same classifier and the same set of

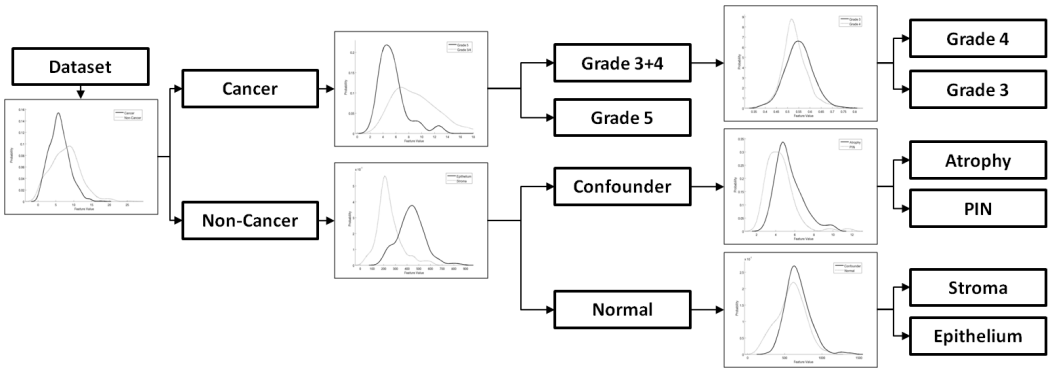


Figure 7.2: Illustration of the cascaded (CAS) approach, where classification is performed between broad class groups on the left and ending with the most granular classes on the right.

features. Assigning multiple decision boundaries can leads to classification errors, particularly when some classes are similar (e.g. different types of cancerous tissue) and others are dissimilar (cancerous and benign tissue). An illustration of this is given in Figure 7.1 (a), where each line represents a probability density function (the likelihood of observing that class given a particular feature value). Clearly, assigning a set of decision boundaries to separate out these classes would lead to suboptimal results. An alternative is the one-versus-all (OVA) approach, where each class is distinguished individually from all non-target classes. Figure 7.1 (b) showcases this approach, where the "Target" class probability is plotted against the "Non-target" class. Since the non-target consists of heterogeneous tissue types, this distribution is multi-modal, and assigning a single classification boundary in this case would lead to misclassification.

A more strategic approach is represented by a cascaded algorithm, as illustrated in Figure 7.2. In this strategy, successive binary classifications are performed where the initial classification is performed by grouping similar classes together according to class similarity and domain knowledge. Each bifurcation in Fig. 7.2 represents a binary classifier that distinguishes dissimilar "class groups," ensuring that the classes within a group are relatively similar. Subsequently, two new binary classifications are used to separate each of the class groups further, again grouping similar sub-classes together. Finally, the most difficult-to-classify images are separated by the final classification. The cascaded approach confers two strong advantages over both OSC and OVA classification: (1) By utilizing multiple independent binary classifiers, we avoid the problem

of having to identify multiple classes at once using the same classifier, and are thus able to tailor the classifiers to the specific task by selecting features and parameters that are optimized for the task at hand. (2) By selecting class groups using domain knowledge, we are able to minimize the class heterogeneity for each classification task, thus avoiding the problems of multi-modal feature distributions as seen in the OVA approach.

In this work, we present a system for classifying regions of interest of prostate tissue into one of seven classes: normal epithelium, normal stroma, Gleason grades 3, 4, and 5 cancer, prostatic intraepithelial neoplasia (PIN), and atrophy. We employ a cascaded approach to this problem by grouping similar classes into class groups and performing binary classification at increasing levels of granularity. We test our classification algorithm by comparing the cascaded approach with two traditional multi-class approaches: the one-shot classification (OSC) approach, where classification algorithms attempt to distinguish all classes simultaneously, and the one-versus-all (OVA) approach, where individual classes are classified independently from all other classes. We show that by incorporating domain knowledge and developing a cascaded approach to classification, we can accurately identify increasingly granular classes.

The layout of the rest of the paper is as follows. Section 7.3 describes the data used in this study as well as the image acquisition and processing techniques. Section 7.4 discusses the extraction and modeling of features from each image, and Section 7.6 describes the cascaded classifier procedure. Section 7.7 discusses the experimental setup, and Section 7.8 provides the results and discussion. Concluding remarks are presented in Section 7.9.

7.3 Image Acquisition and Processing

7.3.1 Prostate Biopsy Tissue Preparation and Digitization

Prostate biopsy samples were acquired from 214 patients at the Department of Surgical Pathology at the University of Pennsylvania in the course of normal clinical treatment. Tissue samples were stained with hematoxylin and eosin (H&E) to highlight nuclear

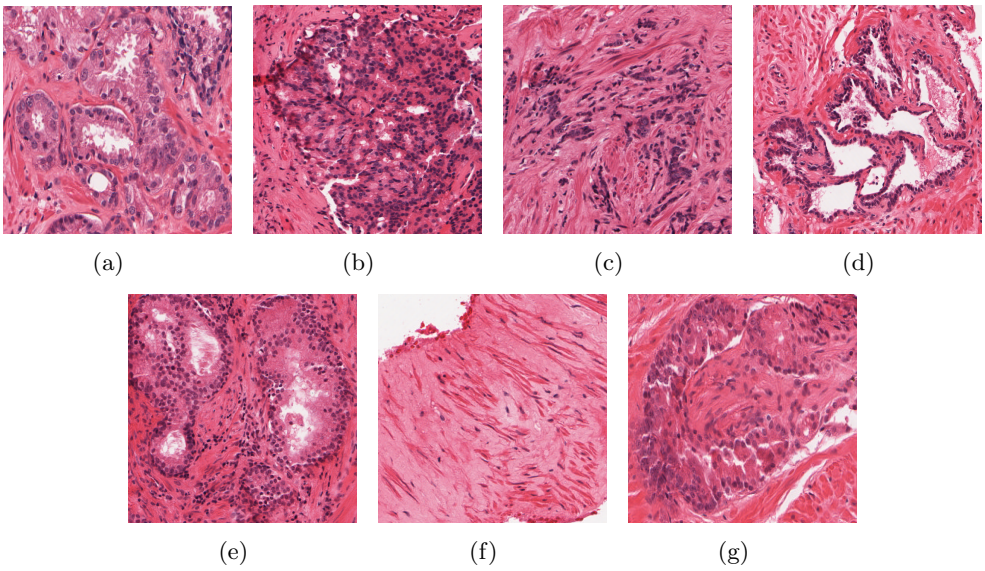


Figure 7.3: Illustration of the different tissue types examined in this study. Shown are ROIs belonging to (a) Gleason grade 3, (b) Gleason grade 4, (c) Gleason grade 5, (d) tissue atrophy, (e) benign epithelium, (f) benign stroma, and (g) prostatic intraepithelial neoplasia.

and cytoplasmic material in the cells. Following fixation, the slides were scanned into a computer workstation at 40x optical magnification using an Aperio whole-slide digital scanner (Aperio, Vista, CA). The acquisition was performed following an automated focusing procedure as per the recommended software settings, and the resulting files were saved as ScanScope Virtual Slide (SVS) file format, which are equivalent to the tagged image file format (TIFF) standard. In terms of pixel size, each image measures from 10,000 to 100,000 pixels in a dimension, depending on the amount of tissue on the slide. Uncompressed images range from 1 gigabyte (GB) to over 20 GB in hard drive space. Images were compressed using the JPEG standard to a quality of 70 (compression ratio of approximately 1:15); at the image magnification that was captured, this compression did not result in a significant loss of quality of the acquired images.

7.3.2 Ground Truth Annotation on Digitized Biopsy Samples

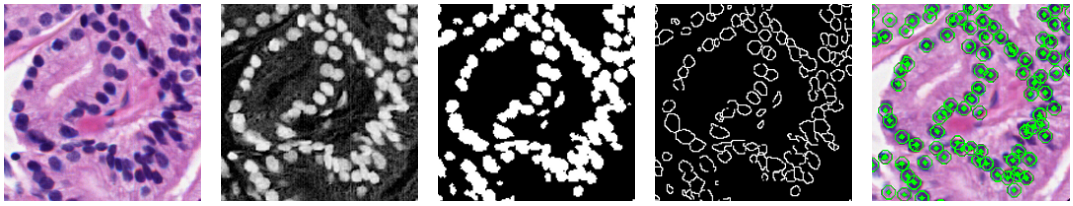
The ScanScope slide viewing software allows the user to manually delineate regions of interest (ROIs) on the image, followed by tagging the regions with metadata such as a description and region size. This information is saved as an extensible markup language (XML) document which can be read and parsed by the program. An expert pathologist

was given instructions to manually annotate ROIs of relatively homogeneous tissue. We denote an ROI as $\mathcal{R} = (R, g)$, where R is a 2D set of pixels $r \in R$ and $g(r)$ is an intensity function that assigns a triplet of intensity values to each pixel (corresponding to the red, green, and blue color channels of the image). The class of $\mathcal{R}$ is denoted as ω_i for $i \in \{1, \cdots, k\}$ classes, and we use the notation $\mathcal{R} \hookrightarrow \omega_i$ to indicate that $\mathcal{R}$ belongs to class ω_i. In this work, $k = 8$, corresponding to one of eight different tissue types (Figure 7.3): Gleason grade 3 (G3), Gleason grade 4 (G4), Gleason grade 5 (G5), benign (normal) epithelium (BE), benign (normal) stroma (BS), perineural invasion (PI), prostatic intraepithelial neoplasia (PIN), and tissue atrophy (AT).

Because there were no size restrictions on the manual annotation, several regions of tissue were labeled at varying sizes and shapes. Thus, ROIs were extracted from the larger biopsy sample by placing a best-fit bounding box on the annotated region. While some amount of tissue may be included that did not belong to the ROI's class, this surrounding tissue was considered part of the microenviromnent of the tissue and was not discarded. Additionally, the area of this extraneous tissue was generally small compared to the rest of the ROI. Finally, it should be noted that the annotation of individual tissue types on pathology is not a common practice within clinical diagnosis and prognosis of prostate biopsy samples. Thus, there are no generally-accepted guidelines for drawing exact boundaries for regions of cancer, PIN, PI, or atrophy; however, the annotating pathologists were only told to try and ensure that the majority of each ROI was from the same tissue class.

7.4 Tissue Architecture Feature Extraction

Tissue classes are identified primarily by the spatial arrangement of the nuclei (Figure 7.3), which indicates the amount of proliferation and distribution of the cells in the tissue. To quantify these characteristics, we extract a number of architectural tissue features, which are based on the number, density, and arrangement of nuclei within the image. The extraction of these features requires the following steps: (1) Detection of cellular nuclei via color deconvolution and watershed segmentation; (2) Construction of graphs using detected nuclear centroids; and (3) Extracting graph-based features to



(a) (b) (c) (d) (e)

Figure 7.4: Overview of automatic nuclei detection. Shown are: (a) the original tissue image, (b), the result of color deconvolution to isolate the nuclear stain, (c) the result of thresholding to get nuclear regions, (d) the result of watershed segmentation of the nuclear boundaries, and (e) the centroids of the detected regions in the watershed image.

quantify nuclear architecture. Each of these steps is described in detail below.

7.4.1 Color Deconvolution for Nuclei Region Detection

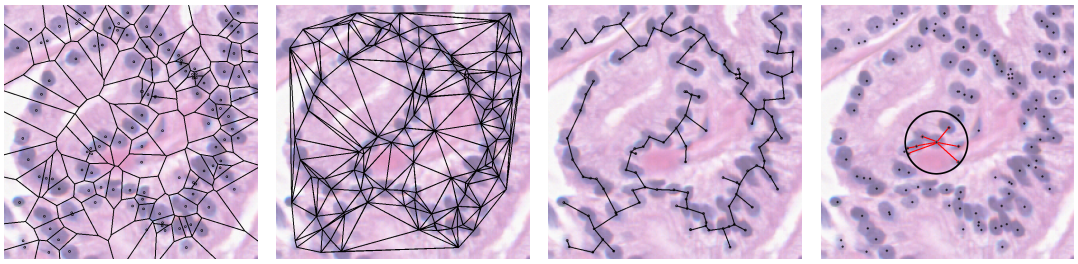
For each $\mathcal{R}$, there exists a set of nuclear centroid pixels $r \in V$. To isolate nuclear regions we use a color deconvolution process detailed in [85]. The optical density (OD) of a material is given by $a = -\log \frac{I}{I_0}$, where I is the intensity of light transmitted through the material (i.e. the light detected by the microscope or scanning hardware), and I_0 is the intensity of light incident on the material [86]. The value of a can be found empirically by measuring the incident and transmitted light for each channel (red, green, and blue) and each stain of an image. We can obtain a normalized three-by-three matrix $\mathbf{M}$ where the rows indicate the materials in the sample (hematoxylin, eosin, and background) and the columns denote the red, green, and blue channels of the image. If we denote by $\mathbf{C}$ the three-element vector representing the amount of each stain at a pixel r, then $g(r) = \mathbf{CM}$ represents the three-element intensity vector at r. We can then solve $\mathbf{C} = g(r)\mathbf{M}^{-1}$ to obtain the amount of each stain present at pixel r [85]. Shown in Figure 7.4(a) is a tissue sample, followed by the result of color deconvolution in Fig. 7.4(b), where the intensity of the pixels is proportional to the amount of hematoxylin stain present in Fig. 7.4(a). Shown is the channel corresponding to the hematoxylin stain (the nuclear material).

7.4.2 Finding Nuclear Centroids Via Watershed Segmentation

The deconvolved image shows the relative amount of stain at each pixel. To obtain the nuclear centroids, we must first binarize the image using Otsu's thresholding method [43] to yield the set of pixels within the nuclear region, $r \in N$. An example of the binarized image is shown in Fig. 7.4(c). The pixels on the boundary of N are denoted as C; these pixels do not include boundaries between overlapping or closely adjacent nuclei. To identify nuclear boundaries within N, we employ a watershed algorithm [87]. The Euclidean distance transform is applied to N to generate a distance image $\mathcal{D} = (R, d)$, where $d(r)$ is the distance from r to the closest point on C. The watershed algorithm is a method of segmenting an object by assuming that high values of d are "valleys" that can be filled, and local maxima are point sources. The points where two pools merge are considered the segmentation of the region. The set of nuclear centroids V is then derived from the geometric center of the segmented nuclei.

7.4.3 Nuclear Architecture Feature Extraction

We denote a graph as $\mathcal{G} = (V, E, W)$, where V are vertices, E are edges, and W are edge weights, proportional to length. The set of vertices, edges, and weights make up a unique graph on $\mathcal{R}$. We construct the following graphs (illustrations are shown in Figure 7.5:)



(a) (b) (c) (d)

Figure 7.5: Examples of the architectural feature extraction performed in this study. Shown are (a) the Voronoi Diagram, (b) Delaunay Triangulation, (c) Minimum Spanning Tree, and (d) nuclear density calculation for the image shown in Figure 7.4 (a).

Voronoi Diagram ($\mathcal{G}_V$)

The Voronoi Diagram partitions $\mathcal{R}$ into a set of polygons with centroids V, where a non-centroid pixel is assigned to the polygon of the closest centroid pixel. This yields a tessellation of the image, as shown in Figure 7.5 (a). Pixels that are equidistant from exactly two centroids make up E (edges of the graph), while pixels equidistant from three or more centroids make up the intersections of multiple edges. The perimeter, area, and chord lengths of each polygon in $\mathcal{G}_V$ are computed, and the average, standard deviation, disorder, and minimum to maximum ratio of each are calculated for a total of 12 Voronoi-based features per $\mathcal{R}$.

Delaunay Triangulation ($\mathcal{G}_D$)

The Delaunay Triangulation is a triangulation of vertices V such that the circumcircle of each triangle contains no other vertices. The Delaunay and Voronoi graphs are dual to each other, meaning that two points are connected in $\mathcal{G}_D$ if and only if their polygons in $\mathcal{G}_V$ share an edge. An example of $\mathcal{G}_D$ is given in Figure 7.5 (b). From this graph, we compute the area and perimeter of each triangle, and the average, standard deviation, disorder, and minimum to maximum ratio of these are calculated to yield 8 Delaunay-based features per $\mathcal{R}$.

Minimum Spanning Tree ($\mathcal{G}_M$)

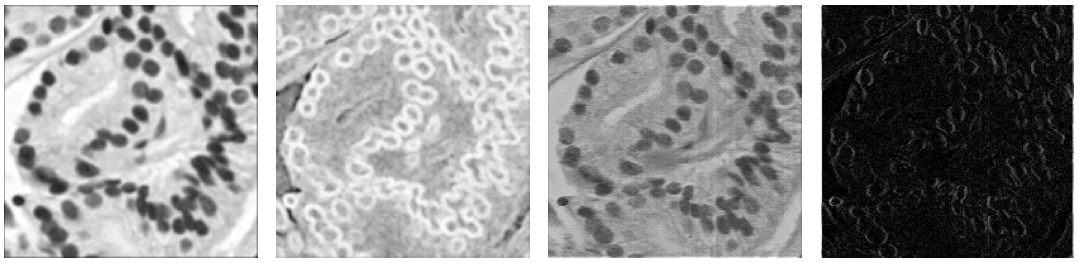
A spanning tree of a set of points V is an undirected, fully connected graph on V. The weight W of the graph is the sum total of all edges E, and the Minimum Spanning Tree is the spanning tree with the lowest overall W. The Minimum Spanning Tree (MST), denoted $\mathcal{G}_M$, is a subgraph of the Delaunay Triangulation. An example of $\mathcal{G}_M$ is given in Figure 7.5 (c). We calculate the average, standard deviation, disorder, and minimum to maximum ratio of the weights W to yield 4 MST-based features per $\mathcal{R}$.

Nuclear Density

Finally, we calculate a set of features that quantify the density of the nuclei without reliance on graph structures. These features are illustrated in Figure 7.5 (d). We compute two sets of nuclear-density based features: (1) We construct a circle around each point in V with a fixed radius (black circle in Fig. 7.5 (d)), and count the number of neighboring points in V that fall within that circle. This is done for radii of 10, 20, 30, 40, and 50 pixels, and for each point in V. The average, standard deviation, and disorder is computed across all points in V to yield 15 features for each $\mathcal{R}$. (2) We calculate the distance from a point in V to the nearest 3, 5, and 7 neighbors (red lines in Fig. 7.5 (d)). This is done for each point in V, and the average, standard deviation, and disorder is computed to yield 9 additional features, for a total of 24 features based on nuclear density.

7.5 Tissue Texture Feature Extraction

The proliferation of nuclei, difference in size and shape of lumen area, and breakdown of typical glandular structure (see Figure 7.3) leads to a change in overall textural characteristics in an ROI. To quantify this change in tissue texture characteristics, we calculate a number of low-level image statistics from each ROI. These statistics can be broadly characterized into three groups: first-order statistics, second-order co-occurrence features, and steerable filter features. Each of these is calculated in a pixel-wise fashion and are computed independently for each of the hue, saturation, and intensity channels of the original scanned image, generating a set of feature images (Figures 7.6 (a)-(d)). The average, standard deviation, and mode of each of these feature images is calculated, yielding a texture feature vector to quantify the image. In total, 468 texture features are calculated in this manner. The details of each feature type are given below.



(a) (b) (c) (d)

Figure 7.6: Examples of the texture feature images generated during feature extraction. Shown are (a) first-order statistics (average intensity), (b) co-occurrence feature values (contrast entropy), and (c), (d) two steerable Gabor filters ($\kappa = 5$, $\theta = \frac{5 \cdot \pi}{6}$) illustrating the real and imaginary response, respectively).

First-order Statistics

We calculate 15 different first-order statistics from each image, including average, median, standard deviation, and range of the image intensities within the sliding neighborhood, as well as the Sobel filters in the vertical, horizontal, and both diagonal axes, 3 Kirsch filter features, gradients in the vertical and horizontal axes, difference of gradients, and diagonal derivative. By calculating these 15 features for each channel in the image, and then calculating the mean, standard deviation, and mode of the feature images, we obtain a total of 135 first-order statistics for $\mathcal{R}$. An example of the average hue feature image is shown in Figure 7.6 (a).

Co-occurrence Features

Co-occurrence features, also referred to as Haralick features [29], are computed by constructing a symmetric 256×256 co-occurrence matrix which describes the frequency with which two different pixel intensities appear together within a fixed neighborhood. The number of rows and columns in the matrix are determined by the maximum possible value in a channel of $\mathcal{R}$; for 8-bit images, this corresponds to $2^8 = 256$. Element (a, b) in the matrix is equal to the number of times pixel value a occurs adjacent to pixel value b in $\mathcal{R}$. From the co-occurrence matrix, a set of 13 Haralick features are calculated: contrast energy, contrast inverse moment, contrast average, contrast variance, contrast entropy, intensity average, intensity variance, intensity entropy, energy, correlation,

entropy, and two measures of information. Extracting these values from each channel and taking the mean, standard deviation, and mode of each feature image yields a total of 117 co-occurrence features. An example of the contrast entropy image is shown in Figure 7.6 (b).

Steerable Filters

A steerable filter refers to a filter which is parameterized by orientation. One such filter is the Gabor filter [40, 31], which is a Gaussian function modulated by a sinusoid. The response of a Gabor filter at a given image coordinate is given as:

$$\mathbf{G}(x, y, \theta, \kappa) = e^{-\frac{1}{2}((\frac{x'}{\sigma_x})^2+(\frac{y'}{\sigma_y})^2)} \cos(2\pi\kappa x'), \tag{7.1}$$

where $x' = x\cos(\theta)+y\sin(\theta)$, $y' = y\cos(\theta)+x\sin(\theta)$, κ is the filter's frequency shift, θ is the filter phase, σ_x and σ_y are the standard deviations along the horizontal and vertical axes. We utilize a filter bank consisting of two different frequency-shift values $\kappa \in \{5, 9\}$ and six orientation parameter values ($\theta = \frac{\epsilon\cdot\pi}{6}$ where $\epsilon \in \{0, 1, \cdots, 5\}$), generating 12 different filters. Each filter yields a real and imaginary response, which is calculated for each of the three channels. An example of two Gabor-filtered images is shown in Figures 7.6 (c) and (d), illustrating the real and imaginary response, respectively, for a filter with $\kappa = 5$ and $\theta = \frac{5\cdot\pi}{6}$. Taking the mean, standard deviation, and mode of each feature image yields a total of 216 steerable filter texture features.

Feature Type	Feature Subtype	Features	Total
Architecture	Voronoi Diagram	Area, chord length, perimeter	12
	Delaunay Triangulation	Area, perimeter	8
	Minimum Spanning Tree	Branch Length	4
	Nuclear Density	Nearest Neighbors, distance to neighbors	24
Texture	First-Order	Statistics, Sobel and Kirsch filters, Gradients	135
	Co-occurrence	Contrast, Intensity, Energy, Correlation, Entropy	117
	Steerable Filter	Frequency and Orientation Parameters	216

Table 7.1: List of the features used in this study, broken into architectural and texture features.

7.6 Cascaded Multi-Class Classification

To classify each ROI, we employ a cascaded approach illustrated in Figure 7.2. The cascaded setup consists of a series of binary classifications, each subdividing the dataset

into increasingly granular class groups. Each bifurcation in Figure 7.2 represents a classification task, amounting to six binary divisions. The motivation for the chosen class groups is based on domain knowledge. We begin by identifying the most broad class group, containing all the samples in the database. For our application, this corresponds to "cancer" vs. "non-cancer". Within the cancer group, we further classify samples into either Gleason grade 5 or a class group containing Gleason grades 3 and 4; this is done because within the cancer group, Gleason grades 3 and 4 are more similar to one another than either is to grade 5. Similarly, we classify non-cancer samples into either the "confounder" class group or the normal class group. Finally, each of the remaining class groups is further classified to obtain the final classification for all samples: Gleason grades 3 and 4 are classified, atrophy and prostatic intraepithelial neoplasia (PIN) are classified, and epithelium and stroma are classified. Each bifurcation is a separate, independent task with its own trained classifier.

For each task, we use a decision tree classifier [45]. Decision trees use a training set of labeled samples to learn a series of rules or "branches" based on feature values. These rules attempt to optimally distinguish between each of the class labels, which are represented by "leaves" at the end of the tree. Classification can then be performed on a testing set, using the features of the testing sample to traverse the tree and arrive at the leaf representing the correct class of the sample. While any classification algorithm may be used in the framework of the cascaded classification, we chose to decision trees for a number of reasons: (1) Decision trees can inherently deal with several classes by creating multiple different class leaves, allowing us to implement the OSC classification strategy directly for comparison. (2) The structure of the tree can be examined to determine which features appear closest to the top of the tree, which are typically the most discriminating features for that classification task. Additionally, these features are selected independently for each of the classification tasks, allowing us to use an optimal set of features for each level of the cascade. In this way we can determine the optimal features to distinguish each class group.

7.7 Experimental Setup

7.7.1 Data Description and Image Details

H&E-stained prostate biopsy tissues are digitized at 40x optical magnification using an Aperio scanner. Regions of interest (ROIs) corresponding to each class of interest are manually delineated by an expert pathologist, with the goal of obtaining relatively homogeneous tissue patches (i.e. patches that express only a single tissue type). Images are analyzed at a downsampled resolution equivalent to 20x optical magnification. Data was collected from a total of 118 patient biopsies. Approximately 1,488 image regions were generated in total, from which 50 samples were chosen at random for each class ω_i per trial (total of 350 samples in each trial).

7.7.2 Classifier Comparison

We performed three-fold cross-validation for ten trials, using two-thirds of the dataset for training and one-third for testing. We tested three different multi-class approaches:

Cascade (CAS): The cascaded strategy is our proposed method, described in Section 7.6.

One-Shot Classification (OSC): In the OSC strategy, the entire dataset is classified into seven classes simultaneously. This is handled implicitly by the decision tree construction, where rule branches terminate at several different class labels.

One-Versus-All (OVA): In the OVA strategy, a binary classifier is used to identify a single target class apart from a single non-target class made up of the remaining classes. Each class is classified independently of the others, meaning that errors in one class do not affect the performance of the others.

We employed the C5.0 implementation of decision trees (DT) [45] to perform the classification, which is an efficient update to the popular C4.5 algorithm. The output of the algorithm consists of the number of samples from each class, and the resulting classification of those samples. This enables us to calculate both the accuracy (ACC) and the positive predictive value (PPV) in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), where:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \qquad \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{7.2}$$

Evaluation is done on a per-class basis, to ensure that comparisons between different classification strategies were standardized.

7.7.3 Feature Ranking

Because of the range of classes being analyzed in this work, we are interested in the discriminating power of the individual features used in this work. The decision tree algorithm identifies the order in which features are used to construct the rule branches; typically the first rule is the most discriminating, and is the most efficient feature to split the dataset. Thus, for each trial, we identify the features that appear within the first five rules of the tree, and the most common feature for each classification task in the cascade are recorded.

7.7.4 Automated Nuclei Detection

We wish to ensure that our automated nuclear detection algorithm (Section 7.4 is performing adequately; however, the purpose of nuclei detection is to extract discriminating features from the tissue samples, as opposed to exact delineation of each nuclear centroid in the image. Therefore, traditional methods of segmentation evaluation (such as percentage overlap, Hausdorff distance, and Dice coefficients) are not appropriate for evaluating this task. To ensure that our feature extraction is performing appropriately, a subset of images from four classes (epithelium, stroma, and Gleason grades 3 and 4) had nuclear centroids manually annotated. We compared the features obtained through our automated detection algorithm, using color deconvolution and watershed segmentation, with the features obtained using manual annotation. Comparison was performed using a Student's T-test to determine how many features had no statistically significant difference between the two sets of feature values.

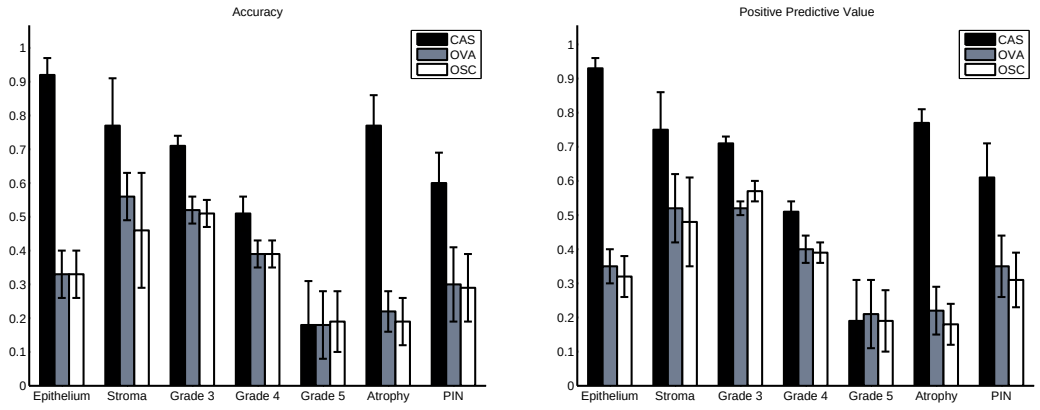
7.8 Results and Discussion

Figure 7.7 illustrates the performance values for each of the classification strategies. Shown are the accuracy (ACC) and PPV values for each of the three classification paradigms, averaged over 20 trials (error bars represent the standard deviation). In the majority of classification tasks, the CAS strategy out-performs both OVA and OSC in terms of ACC and PPV. The average AUC across all classes is 0.64 for CAS, 0.36 for OVA, and 0.34 for OSC, while the average PPV is 0.64 for CAS, 0.37 for OVA, and 0.34 for OSC. Gleason grade 5 tissue is the most difficult to classify in all three strategies; epithelium is most accurately classified in the CAS approach, while stroma and Gleason grade 3 are easiest to classify with OVA and OSC.

For the majority of tissue types there is very little difference between the OVA and OSC approaches. There may be some similarity to the way a sample is classified in both approaches – for example, the tree created in the OSC approach (where all classes are represented in the leaves of the tree) may be a generalized version of the tree in the OVA case; the major difference is in the leaf labels. An in-depth analysis of the trees generated in these cases would indicate whether this was the case. Additionally, it is uncertain whether this result is consistent with other classification algorithms that can perform one-shot classification (e.g. neural networks).

7.8.1 Feature Ranking

The most discriminating features for each class group are shown in the table in Figure 7.8. We find that when attempting to distinguish between tissue types with organized structures, such as “cancer” vs. “non-cancer” or “epithelium” vs. “stroma”, architectural features play a large role. This is largely due to the broad differences in nuclear proliferation between these groups, as well as the regularity of the structures present within each group. However, texture plays a larger role when distinguishing individual Gleason grades, which are structurally similar and have very little regularity in terms of nuclear architecture but have subtle variations that are easily identified by texture features. It should be noted that both texture and architecture did play a role in each



(a) (b)

Figure 7.7: Average performance measures from the three different classification strategies: CAS (our cascaded approach), OSC (one-shot classification), and OVA (one-versus-all classification). Shown are the values for (a) accuracy and (b) positive predictive value, with each group representing a separate tissue class. Error bars represent the standard deviation over 20 trials. The cascaded approach out-performs both OSC and OVA in the majority of tasks, with Grade 5 tissue being the most difficult to classify.

Class	**Highest Weighted Feature**
Cancer vs. Non-cancer	Average Distance to 7 Nearest Neighbors (Architecture)
Grade 5 vs. Grade 3+4	Contrast Average Saturation (Texture)
Grade 3 vs. Grade 4	Standard Deviation Hue Value (Texture)
Confounder vs. Normal	Average Hue Value (Texture)
Atrophy vs. PIN	Average MST Branch Length (Architecture)
Epithelium vs. Stroma	Voronoi Chord Average Length (Architecture)

Figure 7.8: Most discriminating features for each task, as determined by the C5.0 algorithm. For distinguishing highly structured tissue types like cancer vs. non-cancer or epithelium vs. stroma, architecture plays a large role due to the vastly different structures present in each tissue. For more erratic tissue types, such as Gleason grades, texture plays a greater role.

classification task, but the contributions of those features to the overall classification were small compared to the ones listed in Figure 7.8.

7.8.2 Comparison of Manual vs. Automated Nuclei Detection

The results of comparing the feature sets generated via manual and automated nuclei detection are shown in Table 7.2. For each of the four classes with manually-annotated nuclei, we list how many features had $p > 0.05$ and $p > 0.01$, indicating that there was no statistically significant difference between the manually- and automatically-extracted features. We found that at least 9 features (out of the 51 total architectural features) were considered statistically similar in all classes, with Gleason grade 5 and stroma

Class	$p > 0.05$	$p > 0.01$
Epithelium	13	17
Stroma	24	26
Grade 3	6	9
Grade 4	11	15
Grade 5	28	29

Table 7.2: Number of features whose automatically- and manually-extracted features were considered statistically similar by two different criteria ($p > 0.05$ and $p > 0.01$). Stroma and Gleason grade 5 tissue yielded the most similar features, while Gleason grade 3 had the lowest number of similar features.

having the most (over 20) similar features. This is likely due to the lack of complex structure (such as lumen and intra-luminal protein), enabling the automated system to clearly single out the nuclei in the image. In contrast, Gleason grade 3 had the fewest similar features due to the high degree of proliferation of cancer and the presence of gland structures, which leads to a high number of adjacent and overlapping nuclei. These centroids are difficult to correctly identify both manually and algorithmically, so the greatest amount of disagreement is seen in this class. In general, Voronoi features tended to be significantly similar between the two methods, while nuclear density features (which are highly sensitive to false-positive nuclear segmentations) had the least similarity.

Shown in Figure 7.9 are representative graph images obtained via automated nuclei detection (top row, Figs. 7.9 (a)-(d)) and manual annotation (bottom row, Figs. 7.9 (e)-(h)). The tissue region is from the Gleason grade 3 class; even with the dissimilarity between the manually- and automatically-extracted features, there is a qualitative similarity between the graphs generated by both processes; thus, even if the feature values are not statistically significantly different, we can be confident that the automated method can extract the same type of class information as a manual method.

7.9 Concluding Remarks

In this work, we have presented a cascaded multi-class system that incorporates domain knowledge to accurately classify cancer, non-cancer, and confounder tissue classes on H&E stained prostate biopsy samples. By dividing the classification into multiple class groups and performing increasingly granular classifications, we can achieve greater

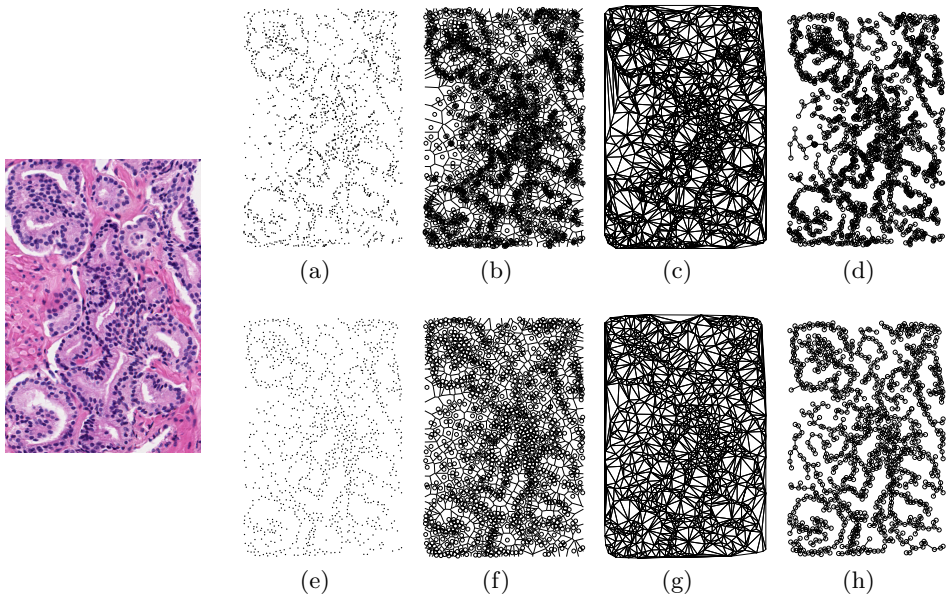


Figure 7.9: Examples of feature images obtained for a Gleason grade 3 image via manual (a)-(b) and automated (e)-(h) nuclear annotations. Shown are the original image at left, followed by the nuclear locations ((a), (e)), Voronoi diagrams ((b), (f)), Delaunay triangulation ((c), (g)), and minimum spanning trees ((d), (f)). Although the automated annotation tends to pick up multiple false positives, the feature values listed in Table 7.2 indicate that the differences are not statistically significant for each image class.

performance when compared to one-shot classification or one-versus-all classification. Our system can be generalized to any multi-class problem that involves classes which can be grouped in a way that maximizes intra-group homogeneity while maximizing inter-group heterogeneity. Additionally, we showed that we can quantify tissue architecture automatically with no statistical difference between our automated and manual nuclei segmentation schemes. Finally, we analyzed the discriminating power of each of our features with respect to each classification task in the cascade, and we found that for class groups with highly structured tissues, architecture plays an important role; however, in cases where tissue types are very similar (i.e. distinguishing Gleason grade), texture is more important to capture the subtle differences in tissue structure.

One limitation of this work is the reliance on domain knowledge rather than image similarity metrics to determine the class groups. Ideally we could calculate an image metric from the training data that would allow us to divide the data into homogeneous groups based on the feature values. Using a proper distance metric to drive the initial design of the system would likely increase the classifier's overall performance. In

addition, we would like to investigate the structure of the trees generated in each of the classification strategies, since the performance of the OSC and OVA strategies was very similar. Alternative classification algorithms capable of performing one-shot classification, such as neural networks, may yield a different result. Likewise, the features selected as highly discriminating may be dependent on the use of decision trees for classification; alternative classification algorithms may separate classes based on different criteria.

Chapter 8

Evaluation of Effects of JPEG2000 Compression on a Computer-Aided Detection System for Prostate Cancer on Digitized Histopathology

8.1 Abstract

A single digital pathology image can occupy over 10 gigabytes of hard disk space, rendering it difficult to store, analyze, and transmit. Though image compression provides a means of reducing the storage requirement, its effects on CAD (and pathologist) performance are not yet clear. In this work we assess the impact of compression on the ability of a CAD system to detect carcinoma of the prostate (CaP) in histological sections. The CAD algorithm proceeds as follows: Glands in the tissue are segmented using a region-growing algorithm. The size of each gland is then extracted and modeled using a mixture of Gamma distributions. A Markov prior (specifically, a probabilistic pairwise Markov model) is employed to encourage nearby glands to share the same class (i.e. cancerous or non-cancerous). Finally, cancerous glands are aggregated into continuous regions using a distance-hull algorithm. We evaluate CAD performance over 12 images compressed at 14 different compression ratios using JPEG2000. Algorithm performance (measured using the under the receiver operating characteristic curves) remains relatively constant for compression ratios up to 1:256. After this point performance degrades precipitously. We also have an expert pathologist view the compressed images and assign a confidence measure as to their diagnostic fidelity.

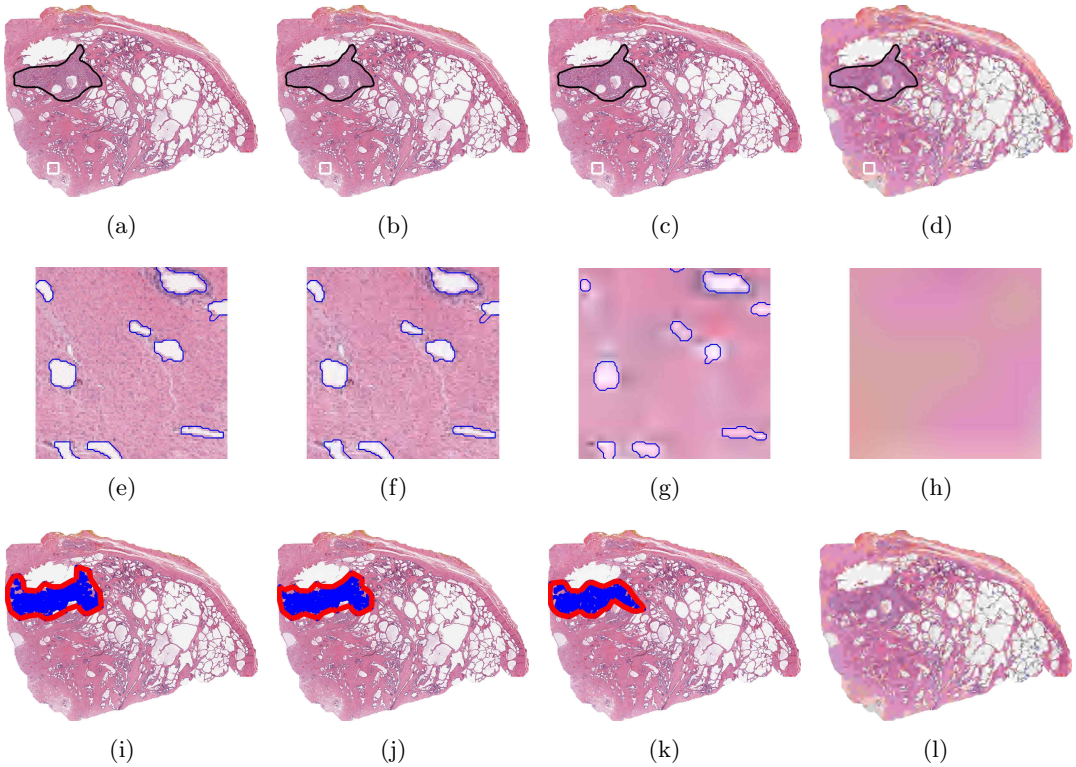


Figure 8.1: JPEG2000 compression on (a) an original histopathology image at (b) 1:16 , (c) 1:256, and (d) 1:4096 compression ratios. Black contours identify the cancer region. The region of interest in a white box is magnified in (e)-(h) to illustrate differences in gland detection and segmentation at different ratios. Shown in (i)-(l) are the results of CAD on each of the compressed images. Results are fairly robust until very high compression ratios. Note that the breakdown of the CAD algorithm occurs at the gland level (h), where detection of glands is impossible.

8.2 Introduction

Digitized images of large tissue samples, such as whole-mount histological sections of the prostate, can constitute more than 10 GB of data. This places a large burden on the computational resources required for storage, transmission, and analysis. On a daily basis a large pathology lab may process hundreds of such studies. This volume of data presents several challenges to digital pathology: 1) Storage becomes prohibitively expensive. 2) Telepathology, the transmission of digital images over computer networks, is untenable. 3) It becomes impossible to employ sophisticated computer-aided diagnosis (CAD) systems. Mitigating these issues requires a method for reducing image file size while retaining diagnostic fidelity [88].

Compression algorithms are a common method for decreasing the storage size of images. The ratio of an uncompressed file size to its compressed size is known as the *compression ratio*. There are two main methods of compression: **lossless**, which are fully reversible but are limited by a low compression ratio, and **lossy**, which achieve high compression ratios at the cost of reduced image quality. In digital pathology, loss of image quality can adversely affect the ability of both a CAD system [89] and a pathologist [90] to perform analysis.

The majority of previous research into the impact of compression of histological images relied on visual quality as measured by a pathologist. For example, Foran, et al. [90] determined which compression ratios were suitable for diagnosis using telepathology. To our knowledge very few papers attempt to quantitatively measure the effects of compression on an automated system: López, et al. [89] found that the differences in nuclei counts as performed by their automated system were not significantly affected by compression ratios of up to 1:46.

In this work we evaluate the impact of JPEG2000 compression on the ability of a computer-aided diagnosis (CAD) system to detect carcinoma of the prostate (CaP) in whole-mount histological sections. We previously developed such a CAD system [91], which proceeds as follows: Step 1) glands are segmented, Step 2) the segmented glands are classified as malignant or benign, and Step 3) the malignant glands are consolidated into continuous regions. The system was shown to detect CaP regions with a sensitivity of 88% and an accompanying false positive rates of the 10% [91]. In the current study, we measure the system performance using 40 whole-mount histology compressed via JPEG2000 at 14 different ratios to determine the robustness of the CAD algorithm to lossy compression. For completeness we also perform a reader study, wherein a pathologist is asked to provide a confidence measure as to the diagnostic fidelity of the compressed images.

8.3 Methodology

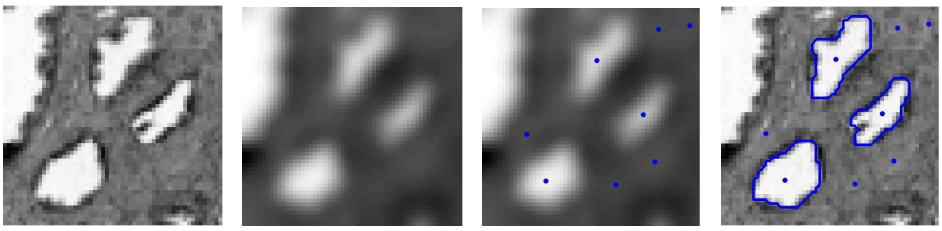
8.3.1 Image Compression Algorithm

The compression scheme used is the JPEG2000 compression standard and coding system, based on the wavelet transform. In JPEG2000 images are first transformed to YUV color space and then convolved with the Cohen-Daubechies-Feauveau discrete wavelet transform. This generates sets of coefficients referred to as sub-bands. Each sub-band represents an approximation of the original image, at a corresponding resolution level, and for each additional sub-band increasing image detail is generated allowing the image to be reconstructed at higher resolution with additional sub-bands. It should be noted that the represented height and width of the image do not change between sub-bands, but the amount of information used to represent each band is increased significantly. Finally, all sub-bands are divided into code blocks of 64-by-64 pixels, which are individually encoded by a three-step Embedded Block Coding with Optimal Truncation (EBCOT) scheme. This scalar-quantization and the code block encoding together determine the amount of compression, known as the compression ratio.

The OpenJPEG implementation of the JPEG2000 standard (http://www.openjpeg.org/) was used to produce nine different compression ratios, each containing a single image consisting of six sub-bands or levels per compressed image. The compression ratios ranged from $1:1$ (no compression) to $1:1024$ (high compression). Examples of compression can be seen in Figure 8.1. The original images are shown in Figure 8.1(a), along with subsequently higher compression ratios (Figures 8.1(b)-8.1(d), respectively).

8.3.2 Cancer Detection and Classification

Figure 8.1(a) illustrates a prostate histological (tissue) section. The pinkish hue results from the H&E staining procedure. The superimposed black line delimits the spatial extent of CaP as determined by a pathologist. The numerous white regions are the gland lumens, i.e. cavities in the prostate through which fluid flows. Our automated system identifies regions of CaP by leveraging two biological properties: 1) cancerous



(a) (b) (c) (d)

Figure 8.2: Overview of the gland detection and segmentation procedure. The luminance channel (a) is convolved with a Gaussian kernel to generate a smoothed image (b). Peaks in this image are used to detect gland centers (c). A region-growing algorithm is used in the unsmoothed image to extract gland size (d). Segmentations with poor average edge strengths are discarded.

glands (and hence their lumens) tend to be smaller in cancerous than benign regions and 2) malignant/benign glands tend to be proximate to other malignant/benign glands.

Gland Segmentation

Figure 8.2 illustrates the gland segmentation procedure. We extract the luminance channel of the digitized section (*CIE Lab* color space), where gland regions appear as contiguous, high intensity pixels circumscribed by sharp boundaries (Figure 8.2(a)). We convolve the image with a Gaussian kernel at multiple scales to generate multiple smoothed images (Figure 8.2(b) illustrates one such image.). The local maxima (i.e. single pixel peaks) are considered to be lumen centers (Figure 8.2(c)), which serve as seeds for a region-growing algorithm (Figure 8.2(d)). We briefly outline this algorithm. First define the following: 1) *current region* (CR) is the set of pixels representing the segmented region in the current step of the algorithm, 2) *current boundary* (CB) is the set of pixels that neighbor CR in an 8-connected sense, but are not in CR, and 3) *internal boundary* (IB) is the subset of pixels in CR that neighbor CB. The growing procedure begins by initializing CR to a seed pixel assumed to lie within the gland. At each iteration CR expands by aggregating the pixel in CB with the greatest intensity. CR and CB are updated, and the process continues. The algorithm terminates when the L_∞ norm from the seed to the next aggregated pixel exceeds a predetermined threshold. That is, the L_∞ norm establishes a square bounding box about the seed; the growing procedure terminates when the algorithm attempts to add a pixel outside this box.

During each iteration the algorithm measures the boundary strength which is defined as the average intensity of the pixels in IB minus the average intensity of the pixels in CB. After the growing procedure terminates, the region with the greatest boundary strength is selected. If the boundary strength is below an empirically-determined signal-to-noise ratio it is discarded.

Gland Classification

Let the set $S = \{1, 2, \ldots, N\}$ reference the N segmented glands in a histological. Each gland has an associated state $X_s \in \Lambda \equiv \{\omega_1, \omega_2\}$, where ω_1 and ω_2 indicate malignancy and benignity, respectively. The random variable $Y_s \in \mathbb{R}$ indicates the area of gland s. Let $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$ refer to all random variables X_s and Y_s in aggregate. The state spaces of $\mathbf{X}$ and $\mathbf{Y}$ are the Cartesian products $\Omega = \Lambda^N$ and $\mathbb{R}^{D \times N}$.

We use maximum *a posteriori* (MAP) estimation to find the optimal $\mathbf{X}$ given the feature vector $\mathbf{Y}$, i.e. we maximize the *a posteriori* probability $P(\mathbf{X}|\mathbf{Y})$. This probability is proportional to the product of the conditional probability $P(\mathbf{Y}|\mathbf{X})$ and the prior distribution $P(\mathbf{X})$. The conditional probability models the area of the glands, as cancerous glands tend to be smaller in size than benign glands [92]. The prior distribution incorporates the biological tendency for cancerous/benign glands to appear near other cancerous/benign glands. More specifically, $P(\mathbf{X})$ is modeled using probabilistic pairwise Markov model (PPMM) [91], a novel Markov prior which is both more flexible and intuitive than typical Markov priors (such as the Potts model). Both the conditional and prior distributions can be learned via training. This approach allows us to generate Receiver Operating Characteristic (ROC) curves for quantitative evaluation as opposed to setting a hard threshold.

Gland Consolidation

Glands determined to be cancerous are consolidated into continuous regions. To perform this consolidation we use a modified form of the convex hull called distance hull or Dhull [91]. Unlike the convex hull, Dhull places a restriction on the maximum distance

between consecutive points on the hull, thus allowing the formation of non-convex boundaries which can better conform to the true CaP regions.

8.4 Experimental Setup and Evaluation

The dataset consists of 40 prostate histology sections stained with hematoxylin and eosin (H&E), obtained from radical prostatectomies at the University of Pennsylvania and Queens University in Canada. Each sample contains regions of CaP ranging in malignancy from Gleason scores six to eight, and is digitized at 1.25x optical magnification (8 μm per pixel) using an Aperio slide scanner. The CaP regions on each digitized sample are manually delineated by a pathologist using a black contour in an image editor.

8.4.1 Experiment 1: Automated Cancer Detection via CAD

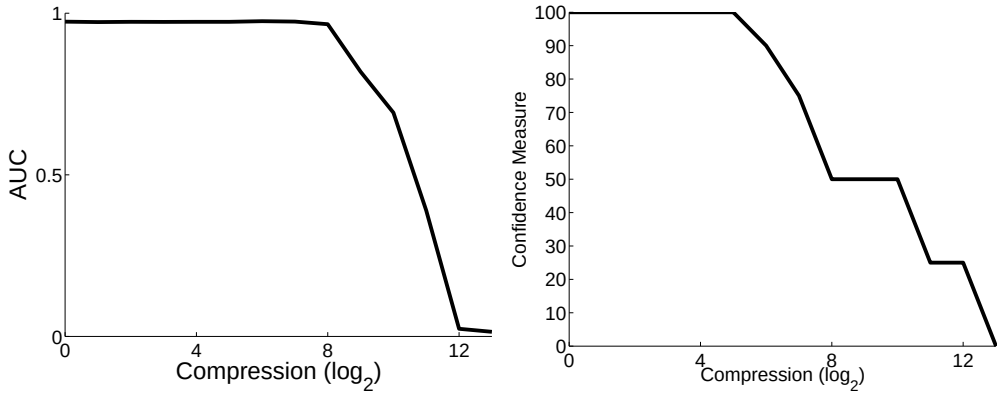
Twenty eight of the histological sections (uncompressed) were used to train the CAD system described in Section 8.3.2. The remaining 12 images were each compressed at 1:1, 1:2, ..., 1:8192, yielding a test set of 168 images. To assess system performance we define the following measure: true positives (TP) indicate the area of the HSs denoted as cancerous by both the pathologist and CAD, and similarly we define true negatives (TN), false positives (FP), and false negatives (FN). From these we obtain two additional measures: the true positive rate TP/(TP+FN) and the false positive rate FP/(TN+FP).

The performance of the CaP detection system with respect to all preceding measures is influenced by the probability that a gland is malignant (or one minus the probability it is benign). This probability can be varied by the user from zero to one, which yields a receiver operator characteristic (ROC) curve. To arrive at a measure that is independent of the prior probability we can calculate the total area under the ROC curve (AUC). Therefore, to evaluate the impact of compression ratio on the performance of the CaP detection system we choose to measure the AUC for each group of 12 images sharing the same compression ratio. This produces 14 total AUCs (one for each compression

ratio).

8.4.2 Experiment 2: Pathologist Reader Visual Inspection

An expert pathologist was instructed to state the confidence in their ability to identify the regions of CaP. The confidence measure ranges from 0 (absence of diagnostic information) to 100 (absolute certainty). To prevent previously-viewed images from influencing subsequent confidence measures, the images were considered serially from the most- to the least-compressed.



(a) (b)

Figure 8.3: (a) Plot of evaluation metric (AUC) as the compression level increases. As compression increases, performance of the CAD algorithm decreases due to a loss of diagnostically useful information. (b) Plot of pathologist confidence in diagnosis as compression level increases. Note that a decrease in pathologist confidence does not indicate incorrect diagnosis, but simply a lack of diagnostically useful information.

8.5 Results and Discussion

8.5.1 Experiment 1: CAD Performance on Compressed Images

AUC vs. Compression Ratio

Quantitative classification results are shown in Figure 8.3(a), with the AUC for each compression ratio plotted as a function of compression ratio. The independent axis is plotted using a log (base 2) scale. For compression ratios up to 1:256 there is very little degradation in classifier performance. At higher compression ratios performance

decreases rapidly; as seen in Figure 8.1(h), the gland detection algorithm can no longer identify the lumens.

Qualitative Evaluation of CaP Regions

Figure 8.1(e) shows a portion of an uncompressed image that contains several glands. Notice that the number and relative sizes and shapes of the gland segmentations are very consistent up to a compression ratio of 1:256 (Figures 8.1(e)-(g)). As the compression ratio reaches increasingly higher levels, details become lost and the algorithm can no longer find the lumen regions (Figure 8.1(h)).

8.5.2 Experiment 2: Reader Inspection of Compressed Images

The reader confidence in classification is plotted in Figure 8.3(b) as a function of compression ratio. The pathologist is quite confident in classifying the cancerous regions in the image until compression ratio reaches around 1:64, at which point confidence decreases to 0%. Note that we are making a distinction between confidence and accuracy: although the pathologist becomes much less confident at ratios exceeding 1:64, this may not necessarily signify a commensurate reduction in detection performance.

8.6 Concluding Remarks

Since digitized histological samples can be several gigabytes in size, image compression is necessary component of digital pathology. Unfortunately, the effects of lossy compression on the analysis of histology images is not well understood. In this paper, we evaluated the impact of image compression with respect to the ability of a CAD algorithm to identify CaP regions on whole-mount histology sections. Specifically, we applied our previously-developed CAD system to images compressed at 14 different compression ratios using JPEG2000. System performance was shown to be very robust for compression ratios up to 1:256. Beyond this level performance dropped off sharply. As can be easily seen in the images in Figure 8.1, this drop-off results from the inability of the CAD system to detect the individual glands. Local high frequency information

is lost at high compression rates, thus we should expect a decline in CAD performance when gland size becomes small in relation to remaining high frequency information (less such information will remain for higher compression rates).

An expert pathologist evaluated the effects of compression on diagnostic image quality. Interestingly, degradation was perceived at compression ratios that did not affect CAD performance. This is not unexpected. Whereas the CAD algorithm only considers the size of the glands, a pathologist interprets additional information such as glandular morphology and the coloring from the H&E stain. Perhaps these attributes degrade more quickly with compression than does glandular area. This suggests that it might be useful to store images at one compression ratio for visual analysis and at another for automated CAD analysis. Additionally, different CAD systems (for the same task) would likely vary in their robustness to compression. For example, those using co-occurrence matrices to extract textural features would likely be very sensitive to the removal of high frequency information. In general, the impact of compression is a function of many factors such as the compression scheme, the general task, and the specific algorithmic implementation. Further research is needed to better understand these dependencies.

Chapter 9

Automated Grading of Breast Cancer Histopathology Using Spectral Clustering With Textural and Architectural Image Features

9.1 Abstract

In this paper we present a novel image-analysis based methodology for automatically distinguishing low, intermediate, and high grades of breast cancer from digitized histopathology. A set of over 3,400 image features, including Haralick textures, Gabor filters, and greylevel statistical features are extracted from each image in a database of 48 hematoxylin and eosin stained breast biopsy tissue studies (30 cancerous and 18 benign images). Spectral clustering (graph embedding) is used to reduce the dimensionality of the feature set, and the resulting eigenvectors are classified using a support vector machine (SVM) trained on a third of the reduced dataset. Individual feature subsets are compared in terms of the accuracy of classification. The system achieves a 95.8% accuracy when discriminating between cancer and non-cancer images, and 93.3% accuracy discriminating between low, intermediate, and high grades of cancerous tissue images. Texture-based characteristics of an image (Gabor filter features) are best able to discriminate between cancerous and non-cancerous tissues, while architectural features best discriminate high and low grades of cancerous tissue due to the discriminability of nuclear architecture. Our methodology is also able to recapitulate the underlying manifold structure on which the different grades of breast cancer lie. The manifold shows a smooth transition from low to intermediate to high grade breast cancer.

9.2 Introduction

Approximately 178,000 new cases of invasive breast cancer are diagnosed and approximately 41,000 women are lost to breast cancer each year in the U.S. (source: *American Cancer Society*). Fortunately, proper screening and diagnostic techniques dramatically increase the survival rate of diagnosed women. The current screening protocol consists of a mammography to identify suspicious regions of the beast, followed by a biopsy of potentially cancerous areas. Biopsy samples are examined under a microscope by a pathologist to determine presence of cancer as well as cancer grade. The malignancy of the disease is commonly determined using the degree of tubule formation, mitotic index, and nuclear pleomorphism, which are assigned a numeric value from 1 (mostly normal) to 3 (mostly abnormal). The Bloom-Richardson (BR) grade is the sum of these three numbers. This scale is commonly used in the United States, and correlates well with disease prognosis [93]. However, it has been shown that there is variability among pathologists when using the system, and differences in diagnosis can lead to suboptimal treatment [94]. Meyer, et al. [95], using the kappa agreement score, found that the BR grading scheme is only moderately reproducible. Due to the importance of histological grading in the diagnosis and proper treatment of patients with breast cancer [96], a quantifiable method of grading breast cancer is desired. Use of Computer-aided Diagnosis (CAD) for breast mammography has been shown to increase radiologist lesion detection sensitivity by as much as 21% [97].

In contrast to CAD for radiology, relatively little work has been done in quantitative image analysis of breast histopathology. Weyn, et al. [83] used wavelets and densiometric features to classify breast tissue nuclei as belonging to high or low grade cancer with an accuracy of 61.52%. Petushi, et al. [98] also found that quantitative histopathological image features contain information that can be used to differentiate between grades of breast cancers.

We have previously developed a CAD methodology for recapitulating prostate cancer Gleason grade on tissue histopathology [3]. In this work, we present a quantitative CAD system for distinguishing between “cancer” and “non-cancer” images of breast

tissue histopathology, as well as between low vs. high BR grade images breast cancer grading from digitized histopathology. An overview of our system is presented in Figure 9.1. Our system calculates over 3,400 textural and architectural features from breast tissue images, which are then used in a spectral clustering (SC) algorithm called graph embedding [99]. The SC algorithm creates a low-dimensional subspace into which the high-dimensional feature data is plotted. In this subspace, the distances between images are preserved such that two images appearing close to one another in the subspace obtained using SC are more similar to one another than two images that appear far apart. This representation allows us to visualize the structure of cancer data and to determine if the BR grade of a cancer image is related to its position in the low-dimensional space. Following spectral clustering, we perform two classification tasks, using a support vector machine (SVM) classifier. Our paper is organized as follows. In Section 2 we describe the methodology of the current work. Section 3 contains the results of our analysis, and in Section 4 we present our concluding remarks.

9.3 Methods

9.3.1 Data Description

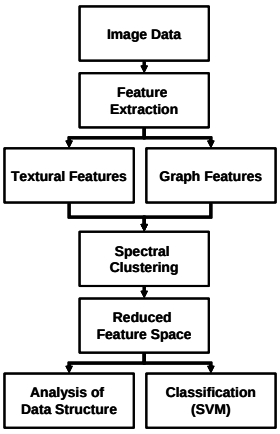


Figure 9.1: System overview.

An overview of our system is presented in Figure 9.1. Glass slides of hematoxylin and eosin stained breast biopsy tissue are scanned into a computer at 40x optical magnification. An expert pathologist labeled regions of tissue within each image according to the Bloom-Richardson (BR) grading scheme [93]. A total of 48 images of tissues taken during routine biopsy were used in this study, comprising 18 benign images and 30 cancerous images. Of the cancer images, 21 were "low grade" cancer (BR grades 5 or 6), and 9 were "high grade" cancer (grades 7 or 8). We denote an image R as $R = (C, f)$, where C is a 2D grid of pixels $c \in C$ and f is a function which assigns a value to c. The pixel value of c is denoted $f(c)$. Each histological image contains m nuclei,

and the pixels corresponding to the centroids of the nuclei are manually labeled. We denote these centroid pixels as $c_n^1, c_n^2, \cdots, c_n^m$.

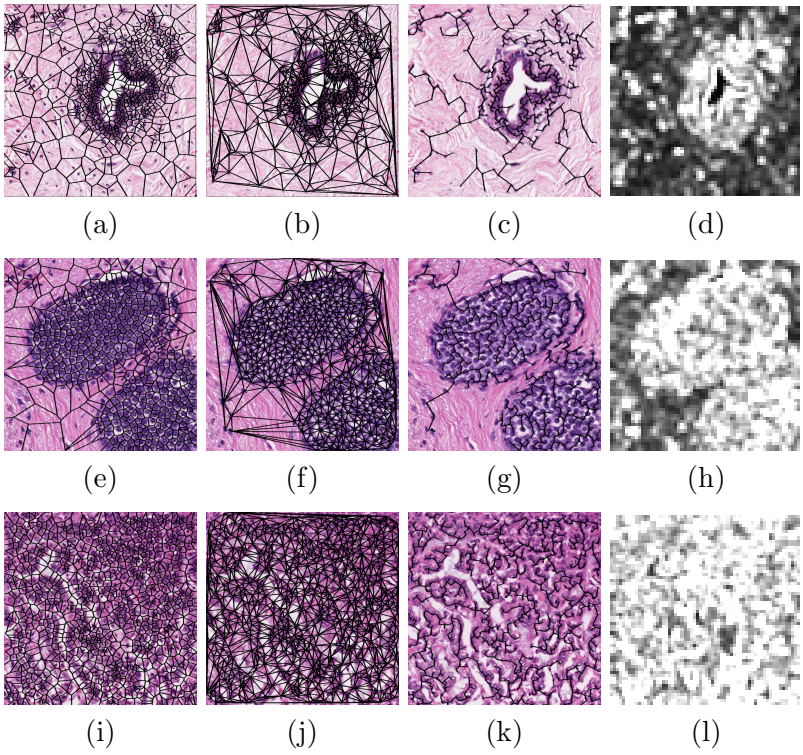


Figure 9.2: Example of the ((a), (e), (i)) Voronoi, ((b), (f), (j)) Delaunay, and ((c), (g), (k)) Minimum Spanning Tree graphs, as well an example of a ((d), (h), (l)) Haralick texture image, calculated for ((a)-(d)) benign tissue, ((e)-(h)) low-grade cancer, and ((i)-(l)) high-grade cancer.

From each image R, we extract a set of graph- and texture-based features to capture the discriminating characteristics of the tissue patterns in each image. A feature vector $\mathbf{f}$ is created for R where each element of $\mathbf{f}$ is a distinct feature value. These values are calculated as described below.

9.3.2 Textural Feature Extraction

The proliferation of nuclei in cancerous tissue suggest that textural characteristics can help discriminate between different grades of breast cancers [83]. The following features are extracted from each of the three channels of an image (hue, saturation, intensity), using three different window sizes (3x3, 5x5, and 7x7 pixels).

Grey Level Features

We calculate 15 gray level features from R as described in [3]. The average, standard deviation, minimum-to-maximum ratio, and mode over all $c \in C$ are then calculated for the values in the feature image to yield a total of 540 gray level feature values for R.

Haralick Features

Second-order co-occurrence texture features are described by 16 Haralick features presented in [29]. We calculate a co-occurrence matrix $Z \in \mathbb{R}^{\mathcal{M}\times\mathcal{M}}$ for image R, which is used to generate 16 Haralick feature images. The average, standard deviation, minimum-to-maximum ratio, and mode of the values in the feature images are calculated to yield 576 Haralick texture feature values for R.

Gabor Filter Features

Steerable Gabor filters respond to a variety of textural differences in an image. A unique filter kernel G is defined by an orientation parameter $\theta \in \{0, \frac{\pi}{8}, \cdots, \frac{7\pi}{8}\}$ and a scale parameter $s \in \{0, 1, \cdots, 7\}$. We construct 64 Gabor filtered images by varying θ and s. The average, standard deviation, minimum-to-maximum ratio, and mode over all $c \in C$ are calculated for each feature image to yield a total of 2,304 Gabor feature values for R.

9.3.3 Graph-based Feature Extraction

The shape and arrangement of nuclei within a histological image region is also related to the cancer progression, and this architecture may be quantified using graph-based techniques [100].

Voronoi Diagram

The Voronoi diagram $\mathbf{V}$ [100] comprises a set of polygons $\mathbf{P} = \{P_1, P_2, \cdots, P_m\}$. Any pixel $c \in R$ is included in polygon P_a if $\mathbf{d}(c, c_n^a) = \min_j\{||c - c_n^j||\}$ where $a, j \in$

$\{1, 2, \cdots, m\}$ and $\mathbf{d}(c, d)$ is the Euclidean distance between any two pixels $c, d \in R$. We calculate the area, perimeter length, and chord length of all $\mathbf{P} \in \mathbf{V}$, and the average, standard deviation, minimum-to-maximum ratio, and disorder [100] are calculated over all $\mathbf{P}$, giving 12 feature values for R.

Delaunay Triangulation

The Delaunay graph $\mathbf{D}$ is constructed such that any two unique nuclear centroids c_n^a and c_n^b, where $a, b \in \{1, 2, \cdots, m\}$, are connected by an edge $E^{a,b}$ if P_a and P_b share a side in $\mathbf{V}$. We calculate the side lengths and areas for all triangles in $\mathbf{D}$, and take the average, standard deviation, minimum-to-maximum ratio, and disorder of these to obtain 8 additional feature values for R.

Minimum Spanning Tree

Given a connected, undirected graph, a spanning tree $\mathbf{S}$ of that graph is a subgraph that connects all vertices. A single graph can have many different $\mathbf{S}$. Weights $\omega_{\mathbf{S}}^E$ are assigned to each edge E in each $\mathbf{S}$ based on the length of E in $\mathbf{S}$. The sum of all weights $\omega_{\mathbf{S}}^E$ in each $\mathbf{S}$ is determined to give the weight $\widehat{\omega}_{\mathbf{S}}$ assigned to each $\mathbf{S}$. The minimum spanning tree (MST) denoted by $\mathbf{S}'$ has a weight $\widehat{\omega}'_{\mathbf{S}}$ less than or equal to $\widehat{\omega}_{\mathbf{S}}$ for every other spanning tree $\mathbf{S}$. We calculate the average, standard deviation, minimum-to-maximum ratio, and disorder of the branch lengths in $\mathbf{S}'$ to obtain 4 additional features for R.

Nuclear Features

Nuclear density $\Pi^{\mathcal{D}}$ is computed as $\Pi^{\mathcal{D}} = \frac{m}{|R|}$, where $|R|$ is the cardinality of R. For each nuclear centroid c_n^a, $N(\mu, c_n^a)$ is the set of pixels $c \in R$ contained within a circle with its center at c_n^a and radius μ. We compute the number of c_n^j, $j \neq a$, $j, a \in \{1, 2, \cdots, m\}$ which are in set $N(\mu, c_n^a)$ for $\mu \in \{10, 20 \cdots, 50\}$. We also compute the μ required to obtain $N(\mu, c_n^a) \in \{3, 5, 7\}$. The average, standard deviation, and disorder of these values for all c_n^j in R is calculated to yield 24 additional features for R.

Figure 9.2 illustrates these graphs for normal benign tissue (Figure 9.2 (a)-(d)), low-grade cancer (Figure 9.2 (e)-(h)), and high-grade cancer (Figure 9.2 (i)-(l)).

9.3.4 Spectral Clustering

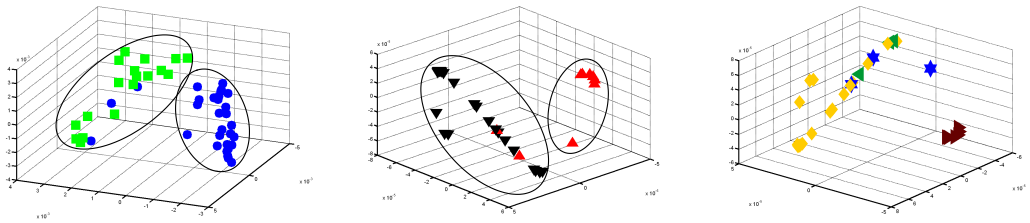
Spectral clustering (SC) algorithms reduce the dimensionality of a data set from M to M' dimensions, where $M' << M$. The motivations for using spectral clustering are twofold: (1) to visualize the high-dimensional manifold of the data in a linear low-dimensional space, and (2) to avoid the *curse of dimensionality*, a phenomenon that decreases classification accuracy when the number of dimensions greatly exceeds the number of classification objects. In the case where $M' = 3$, we can plot the images as points in $\mathbb{R}^3$ such that their linear distance is a measure of their similarity in high-dimensional space. We used graph embedding [99], a nonlinear technique which seeks to find an optimal projection of the data. We construct a confusion matrix $\mathcal{Y}$ to describe the pairwise similarity between the database of images:

$$\mathcal{Y}(p,q) = e^{-||\mathbf{f}_p - \mathbf{f}_q||} \in \mathbb{R}^{\mathbf{N}\times\mathbf{N}}, \tag{9.1}$$

where $\mathbf{f}_p$ and $\mathbf{f}_q$ are the feature vectors computed for any two images R_p and R_q, respectively, where $p, q \in \{1, 2, \cdots, \mathbf{N}\}$ and where $\mathbf{N}$ is the total number of images in the data set. The embedding vector $\mathcal{X}$ is obtained from the maximization of the function:

$$\mathcal{E}_{\mathcal{Y}}(\mathcal{X}) = 2\eta \frac{\mathcal{X}^{\mathbf{T}}(\mathcal{B} - \mathcal{Y})\mathcal{X}}{\mathcal{X}^{\mathbf{T}}\mathcal{B}\mathcal{X}}, \tag{9.2}$$

where $\mathcal{B}(p,p) = \sum_q \mathcal{Y}(p,q)$ and $\eta = |\mathbf{N}| - 1$. The M'-dimensional embedding space is defined by the eigenvectors corresponding to the smallest M' eigenvalues of $(\mathcal{B} - \mathcal{Y})\mathcal{X} = \lambda\mathcal{B}\mathcal{X}$. For image R defined by feature vector $\mathbf{f}$, the embedding $\mathcal{X}(R)$ contains the coordinates of R in the embedding space and is given as $\mathcal{X}(R) = [w_z(R)|z \in \{1, 2, \cdots, M'\}]$, where $w_z(R)$ are the z eigenvalues associated with $\mathcal{X}(R)$.



(a) (b) (c)

Figure 9.3: Graph Embedding results for (a) cancerous (blue circles) vs. non-cancerous images (green squares), (b) high-grade (red up-triangles) vs. low-grade images (black down-triangles), and (c) individual grades of images: Grade 5 (orange diamonds), Grade 6 (green left-triangles), Grade 7 (blue stars), and Grade 8 (maroon right-arrows). Note that the manifold in (b) and (c) is the same; only the view and the labels on the data have been changed. The manifold structure in (c) reveals a smooth transition in BR grade from low-, to intermediate-, to high-grade cancer.

9.4 Results and Discussion

Following SC, we employ support vector machines (SVMs) to classify the images in the database. In this study we perform two classification tasks: (1) distinguishing cancer from non-cancer images, and (2) distinguishing high from low grades of cancer. For each task we divide the feature set into groups to determine if any subset of features outperforms others in either of the two tasks. The 10 feature groups are listed in the first column in Table 9.1. Feature subsets are reduced with SC and used for classification. For each classification task, a third of the dataset is randomly selected for training, while all remaining images are used for testing. We also plot the low-dimensional embeddings to analyze the spatial relationships between images in the reduced embedding space provided by SC.

9.4.1 Quantitative Results

Classification accuracies are given in Table 9.1. The highest accuracy in distinguishing cancerous from non-cancerous images is 95.8%, obtained when using only Gabor filter features. The highest classification accuracy obtained when distinguishing high from low grade cancer images is 93.3%, obtained using all of the architectural features: Voronoi, Delaunay, MST, and nuclear features.

Feature Subtype	**Classification Task**	
	Cancer vs. Benign	**Low vs. High Grade**
All Features (3,468)	0.667	0.700
All Textural (3,420)	0.667	0.733
All Architectural (51)	0.771	**0.933**
Gabor (2,304)	**0.958**	0.700
Grey Level (540)	0.938	0.700
Haralick (576)	0.625	0.767
Voronoi (12)	0.792	0.900
Delaunay (8)	0.854	0.900
MST (4)	0.938	0.900
Nuclear (27)	0.729	0.900

Table 9.1: Classification accuracy using different feature subsets for cancer vs. non-cancer images and high vs. low grade images.

9.4.2 Qualitative Results

The scatter plots in Figure 9.3 show the data in the low-dimensional space obtained through spectral clustering. The axes of the plots correspond to the three dominant eigenvectors found by SC. Figure 9.3 (a) shows cancerous images (blue circles) versus non-cancer images (green squares). Figure 9.3 (b) shows high-grade cancer images (red up-triangles) versus low-grade cancer images (black down-triangles). Black ellipses denote class clusters in both figures. Figure 9.3 (c) shows a rotated view of (b), with labels altered to show BR grades of cancer: the orange diamonds are grade 5, green left-triangles are grade 6, blue stars are grade 7, and maroon right-arrows are grade 8. There is a transition from the low grade images on the left of the plot through intermediate grades (6 and 7) in the middle, ending with high grade images on the right. This suggests that the structure of the data reflects the biology of the images, and that by analyzing this structure in a low-dimensional subspace, we can appreciate the relationship between the underlying biology and the image features that are calculated from the image.

9.5 Concluding Remarks

In this paper, we present an automated system for quantitative histopathological analysis of breast tissue images. We have described a set of image features to extract information used to distinguish different tissue patterns. Quantitative classification

results indicate that our system is capable of accurately distinguishing between cancer and benign images, as well as between images of high and low grades of cancer. Qualitative results from graph embedding illustrate that our methodology is able to recapitulate the underlying manifold structure on which the different grades of breast cancer lie. The manifold shows a smooth transition from low to intermediate to high grade breast cancer. The features we calculate represent informative structure in the high-dimensional space, and an image's position on the structure is linked to the underlying tissue biology. With a densely populated manifold, we can begin to investigate if there is a connection between an image's location on the high-dimensional manifold and the biological grade of the cancer growth, and whether images that appear on different parts of the manifold are biologically different from the rest of the data. In future work we hope to obtain more data to confirm our findings.

References

[1] S. Doyle, A. Madabhushi, J. Tomaszeweski, and M. Feldman. A boosting cascade for automated detection of prostate cancer from digitized histology. In *Medical Image Computing and Computer-Assisted Intervention MICCAI 2006*, pages 504–511, 2006.

[2] S. Doyle, C. Rodriguez, A. Madabhushi, J. Tomaszeweski, and M. Feldman. Detecting prostatic adenocarcinoma from digitized histology using a multi-scale hierarchical classification approach. pages 4759–4762, 30 2006-Sept. 3 2006.

[3] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, J. Tomasezweski, and M. Feldman. Automated grading of prostate cancer using architectural and textural image features. In *Proc. 4th IEEE ISBI 2007*, pages 1284–1287, 2007.

[4] S. Doyle et al. Using manifold learning for content-based image retrieval of prostate histopathology. In *Workshop on CBIR for Biomedical Image Archives, (MICCAI)*, New York, New York, 2007.

[5] 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 *ISBI*, pages 1284–1287, 12–15 April 2007.

[6] S. Doyle et al. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. *ISBI 2008. 5th IEEE Int. Symp.*, pages 496–499, 2008.

[7] S. Doyle et al. A class balanced active learning scheme that accounts for minority class problems: Applications to histopathology. *OPTIMHisE Workshiop (in conjunction with MICCAI)*, pages 19–30, 2009.

[8] S. Doyle et al. A boosted bayesian multi-resolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE Trans. on Biomed. Eng. (In Press)*, 2010.

[9] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 148–156, 1996.

[10] J. Chappelow, N. Bloch, N. Rofsky, E. Genega, R. Lenkinski, W. DeWolf, S. Viswanath, and A. Madabhushi. Collinarus: Collection of image-derived nonlinear attributes for registration using splines. *SPIE Medical Imaging*, 7260, 2009.

[11] D.F. Gleason. Classification of prostatic carcinomas. *Cancer Chemotherapy Reports*, 50(3):125–128, 1966.

[12] B.R. Matlaga, L.A. Eskew, and D.L. McCullough. Prostate biopsy: indications and technique. *Journal of Urology*, 169(1):12–19, 2003.

[13] H.G. Welch, E.S. Fisher, D.J. Gottlieb, and M.J. Barry. Detection of prostate cancer via biopsy in the medicare-seer population during the psa era. *Journal of the National Cancer Institute*, 99(18):1395–1400, 2007.

[14] D. Bostwick and I. Meiers. Prostate biopsy and optimization of cancer yield. *European Urology*, 49(3):415–417, 2006.

[15] W. C. Allsbrook, K. A. Mangold, M. H. Johnson, R. B. Lane, C. G. Lane, and J. I. Epstein. Interobserver reproducibility of gleason grading of prostatic carcinoma: general pathologist. *Human Pathology*, 32(1):81–88, 2001.

[16] J.I. Epstein, P.C. Walsh, and F. Sanfilippo. Clinical and cost impact of second-opinion pathology. review of prostate biopsies prior to radical prostatectomy. *American Journal of Surgical Pathology*, 20(7):851–857, 1996.

[17] G. Alexe, J. Monaco, S. Doyle, A. Basavanhally, A. Reddy, M. Seiler, S. Ganesan, G. Bhanot, and A. Madabhushi. Towards improved cancer diagnosis and prognosis using analysis of gene expression data and computer aided imaging. *Experimental Biology and Medicine*, 234:860–879, 2009.

[18] J. Kong, O. Sertel, H. Shimada, K.L. Boyer, J.H. Saltz, and M.N. Gurcan. Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation. *Pattern Recognition*, 42:1080–1092, 2009.

[19] A. Basavanhally, S. Ganesan, S. Agner, J. Monaco, M. Feldman, J. Tomaszewski, G. Bhanot, and A. Madabhushi. Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. *Biomedical Engineering, IEEE Transactions on*, 57(3):642–653, 2010.

[20] D. Glotsos, I.Kalatzis, P. Spyridonos, S. Kostopoulos, A. Daskalakis, E. Athanasiadis, P. Ravazoula, G. Nikiforidis, and D. Cavourasa. Improving accuracy in astrocytomas grading by integrating a robust least squares mapping driven support vector machine classifier into a two level grade classification scheme. *Computer Methods and Programs in Biomedicine*, 90(3):251–261, 2008.

[21] A.W. Wetzel, R. Crowley, S.J. Kim, R. Dawson, L. Zheng, Y.M. Joo andY. Yagi, J. Gilbertson, C. Gadd, D.W. Deerfield, and M.J. Becich. Evaluation of prostate tumor grades by content based image retrieval. *Proc. of SPIE*, 3584:244–252, 1999.

[22] A.N. Esgiar, R.N.G. Naguib, B.S. Sharif, M.K. Bennett, and A. Murray. Fractal analysis in the detection of colonic cancer images. *IEEE Transactions on Information Technology in Biomedicine*, 6(1):54–58, 2002.

[23] A. Tabesh, M. Teverovskiy, H.Y. Pang, V.P. Kumarand D. Verbel, A. Kotsianti, and O. Saidi. Multifeature prostate cancer diagnosis and gleason grading of histologicalimages. *IEEE Transactions on Medical Imaging*, 26(10):1366–1378, 2007.

[24] J. Diamond, N.H. Anderson, P.H. Bartels, R. Montironi, and P.W. Hamilton. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human Pathology*, 35(9):1121–1131, 2004.

[25] R. Farjam, H. Soltanian-Zadeh, K. Jafari-Khouzani, and R.A. Zoroofi. An image analysis approach for automatic malignancy determination of prostate pathological images. *Cytometry Part B (Clinical Cytometry)*, 72(B):227–240, 2007.

[26] P. Huang and C. Lee. Automatic classification for pathological prostate images based on fractal analysis. *IEEE Transactions on Medical Imaging*, 28(7):1037–1050, 2009.

[27] K. Jafari-Khouzani and H. Soltanian-Zadeh. Multiwavelet grading of pathological images of prostate. *IEEE Transactions on Biomedical Engineering*, 50(6):697–704, 2003.

[28] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[29] R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621, 1973.

[30] D. Gabor. Theory of communication. *Proceedings of the Institution of Electrical Engineering*, 93(26):429–457, 1946.

[31] B.S. Manjunath and W.Y. Ma. Texture features for browsing and retrieval of image data. *Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[32] A. Sboner, C. Eccher, E. Blanzieri, P. Bauer, M. Cristofolini, G. Zumiani, and S. Forti. A multiple classifier system for early melanoma diagnosis. *Artificial Intelligence in Medicine*, 27:29–44, 2003.

[33] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M.K. Chung, and S.C. Johnson. Spatially augmented lpboosting for ad classification with evaluations on the adni dataset. *Neuroimage*, 48:138–149, 2009.

[34] A. Madabhushi, J. Shi, M. Feldman, M. Rosen, and J. Tomaszewski. Comparing ensembles of learners: Detecting prostate cancer from high resolution mri. *Computer Vision Methods in Medical Image Analysis*, LNCS 4241:25–36, 2006.

[35] R.A. Ochs, J.G. Goldin, F. Abtin, H.J. Kim, K. Brown, P. Batra, D. Roback, M.F. McNitt-Gray, and M.S. Brown. Automated classification of lung bronchovascular anatomy in ct using adaboost. *Medical Image Analysis*, 11:315–324, 2007.

[36] Leo Brieman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[37] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[38] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communication*, 31(4):532–540, 1983.

[39] H.D. Cheng, X.H. Jiang, Y. Sun, and J. Wang. Color image segmentation: advances and prospects. *Pattern Recognition*, 34(12):2259–2281, 2001.

[40] A. Jain and F. Farrokhnia. Unsupervised texture segmentation using gabor filters. *IEEE Int. Conf. on Sys., Man, Cyber.*, pages 14–19, 1990.

[41] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, 2001.

[42] N. Borley and M.R. Feneley. Prostate cancer: diagnosis and staging. *Asian journal of Andrology*, 11:74–80, 2009.

[43] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9:62–66, 1979.

[44] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006.

[45] J.R. Quinlan. Decision trees and decision-making. *IEEE Transactions on Systems, Man and Cybernetics*, 20(2):339–346, 1990.

[46] C. Van der Walt and E. Barnard. Data characteristics that determine classifier performance. *17th Annual Symposium of the Pattern Recognition Association of South Africa*, pages 6–12, 2006.

[47] M. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009.

[48] A. Madabhushi et al. Review: Integrated diagnostics: A conceptual framework with examples. *Clin. Chem. and Lab. Med.*, pages 989–998, 2010.

[49] J.P. Monaco, J.E. Tomaszewski, M.D. Feldman, I. Hagemann, M. Moradi, P. Mousavi, A. Boag, C. Davidson, P. Abolmaesumi, and A. Madabhushi. High-throughput detection of prostate cancer in histological sections using probabilistic pairwise markov models. *Medical Image Analysis*, 14(4):617–629, 2010.

[50] H. Fatakdawala, J. Xu, A. Basavanhally, G. Bhanot, S. Ganesan, M. Feldman, J. Tomaszewski, and A. Madabhushi. Expectation maximization driven geodesic active contour with overlap resolution (emagacor): Application to lymphocyte segmentation on breast cancer histopathology. *Biomedical Engineering, IEEE Transactions on*, 57(7):1676–1689, 2010.

[51] H S Seung, M Opper, and H Smopolinsky. Query by committee. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, New York, NY, USA, 1992. ACM.

[52] J Schmidhuber, J Storck, and S Hochreiter. Reinforcement driven information acquisition in non-deterministic environments. Technical report, Fakultät für Informatik, Technische Universität München, 1995.

[53] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994. 10.1007/BF00993277.

[54] S Tong and D Koller. Active learining for structure in bayesian networks. In *IJCAI'01: Proceedings of the 17th International Joint Conference on Artificial Intelligence*, volume 17, pages 863–869, 2001.

[55] David Cohn, Zoubin Ghahramani, and M.I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[56] M S Lee, J K Rhee, B H Kim, and B T Zhang. Aesnb: Active example selection with naïve bayes classifier for learning from imbalanced biomedical data. In *BIBE '09: Proceedings of the 2009 Ninth IEEE International Conference on Bioinformatics and Bioengineering*, pages 15–21, 2009.

[57] S. Veeramachaneni, F. Demichelis, E. Olivetti, and P. Avesani. Active sampling for knowledge discovery from biomedical data. In Alpio Jorge, Lus Torgo, Pavel Brazdil, Rui Camacho, and Joo Gama, editors, *Knowledge Discovery in Databases: PKDD 2005*, volume 3721 of *Lecture Notes in Computer Science*, pages 343–354. Springer Berlin / Heidelberg, 2005.

[58] G M Weiss and F Provost. The effect of class distribution on classifier learning: An empirical study. Technical report, Rutgers University, 2001. Department of Computer Science.

[59] N Japkowicz and S Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

[60] N V Chawla, K W Bowyer, L O Hall, and W P Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[61] J. Zhu and E. Hovy. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 783–790, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[62] G. Batista, A. Carvalho, and M. Monard. Applying one-sided selection to unbalanced datasets. In Osvaldo Cairo, L. Sucar, and Francisco Cantu, editors, *MICAI 2000: Advances in Artificial Intelligence*, volume 1793 of *Lecture Notes in Computer Science*, pages 315–325. Springer Berlin / Heidelberg, 2000.

[63] K. Yang, Z. Cai, J. Li, and G. Lin. A stable gene selection in microarray data analysis. *BMC Bioinformatics*, 7(1):228, 2006.

[64] E. Cosatto, M. Miller, H.P. Graf, and J.S. Meyer. Grading nuclear pleomorphism on histological micrographs. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1 –4, 8-11 2008.

[65] G. Begelman, M. Pechuk, E. Rivlin, and E. Sabo. System for computer-aided multiresolution microscopic pathology diagnostics. In *Proc. IEEE International Conference on Computer Vision Systems ICVS '06*, pages 16–16, 04–07 Jan. 2006.

[66] M. Bloodgood and K. Vijay-Shanker. Taking into account the differences between actively and passively acquired data: the case of active learning with support vector machines for imbalanced datasets. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 137–140, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[67] L. Brieman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[68] Z. Tu. Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. *ICCV*, 2:1589–1596, 2005.

[69] M. Li and I. K. Sethi. Confidence-based active learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1251–61, 2006. Journal Article United States.

[70] ACS. *Cancer Facts and Figures 2010.* Atlanta: American Cancer Society, 2010.

[71] S. Viswanath, B. Bloch, M. Rosen, J. Chappelow, R. Toth, R. Lenkinski N. Rofsky, E. Genega, A. Kalyanpur, and A. Madabhushi. Integrating structural and functional imaging for computer assisted detection of prostate cancer on multiprotocol in vivo 3 tesla mri. *SPIE Medical Imaging*, 7260, 2009.

[72] Jonathan I Epstein, William C Allsbrook, Mahul B Amin, Lars L Egevad, and I. S. U. P. Grading Committee. The 2005 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. *Am J Surg Pathol*, 29(9):1228–1242, Sep 2005.

[73] M. Roula, J. Diamond, A. Bouridane, P. Miller, and A. Amira. A multispectral computer vision system for automatic grading of prostatic neoplasia. In *Proc. IEEE International Symposium on Biomedical Imaging*, pages 193–196, 7–10 July 2002.

[74] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi. A boosted bayesian multi-resolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE Transactions on Biomedical Engineering (Accepted)*, 2010.

[75] Chunming Li, Chenyang Xu, Changfeng Gui, and M.D. Fox. Level set evolution without re-initialization: a new variational formulation. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2005*, volume 1, pages 430–436, 20–25 June 2005.

[76] Reza Farjam, Hamid Soltanian-Zadeh, Kourosh Jafari-Khouzani, and Reza A Zoroofi. An image analysis approach for automatic malignancy determination of prostate pathological images. *Cytometry B Clin Cytom*, 72(4):227–240, Jul 2007.

[77] A. Madabhushi and D. N. Metaxas. Combining low-, high-level and empirical domain knowledge for automated segmentation of ultrasonic breast lesions. *IEEE Trans Med Imaging*, 22(2):155–169, Feb 2003.

[78] Henning Muller et al. A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *Int. J. of Med. Inf.*, 73(1):1–23, 2003.

[79] Lei Zheng, A.W. Wetzel, J. Gilbertson, and M.J. Becich. Design and analysis of a content-based pathology image retrieval system. 7(4):249–255, Dec. 2003.

[80] M.H.C. Law and A.K. Jain. Incremental nonlinear dimensionality reduction by manifold learning. *IEEE Trans. on Patt. Analy. and Mach. Learn.*, 28(3):377–391, 2006.

[81] Jun Kong, O. Sertel, H. Shimada, K. Boyer, J. Saltz, and M. Gurcan. Computer-aided grading of neuroblastic differentiation: Multi-resolution and multi-classifier approach. In *Proc. IEEE International Conference on Image Processing ICIP 2007*, volume 5, pages V–525–V–528, Sept. 16 2007–Oct. 19 2007.

[82] S. Petushi, C. Katsinis, C. Coward, F. Garcia, and A. Tozeren. Automated identification of microstructures on histology slides. In *Proc. IEEE International Symposium on Biomedical Imaging: Macro to Nano*, pages 424–427, 15–18 April 2004.

[83] B. Weyn et al. Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. *Cytometry*, 33(1):32–40, Sep 1998.

[84] D.G. Bostwick and J. Qian. High-grade prostatic intraepithelial neoplasia. *Modern Pathology*, 17:360379, 2004.

[85] A.C. Ruifrok and D.A. Johnston. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23:291–299, 2001.

[86] IUPAC. *Compendium of Chemical Terminology*. Blackwell Science, 1997.

[87] F. Meyer. Topographic distance and watershed lines. *Signal Processing*, 38(1):113–125, 1994.

[88] M Gao, P Bridgman, and S Kumar. Computer aided prostate cancer diagnosis using image enhancement and jpeg2000. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 5203, pages 323–334, 2003.

[89] Carlos Lpez, Marylne Lejeune, Patricia Escriv, Ramn Bosch, Maria Teresa Salvad, Lluis E Pons, Jordi Baucells, Xavier Cugat, Toms Alvaro, and Joaqun Jan. Effects of image compression on automatic count of immunohistochemically stained nuclei in digital images. *J Am Med Inform Assoc*, 15(6):794–798, 2008.

[90] D. J. Foran, P. P. Meer, T. Papathomas, and I. Marsic. Compression guidelines for diagnostic telepathology. *IEEE Trans Inf Technol Biomed*, 1(1):55–60, Mar 1997.

[91] J. Monaco, S. Viswanath, and A. Madabhushi. Weighted iterated conditional modes for random fields: Application to prostate cancer detection. In *Workshop on Probabilistic Models for Medical Image Analysis (in conjunction with MICCAI)*, 2009.

[92] V. Kumar, A. Abbas, and N. Fausto. *Robbins and Cotran Pathologic Basis of Disease*. Saunders, 2004.

[93] H. J. Bloom and W. W. Richardson. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *Br J Cancer*, 11(3):359–377, Sep 1957.

[94] L. W. Dalton et al. Histologic grading of breast cancer: linkage of patient outcome with level of pathologist agreement. *Mod Pathol*, 13(7):730–735, Jul 2000.

[95] J. S. Meyer et al. Breast carcinoma malignancy grading by bloom-richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. *Mod Pathol*, 18(8):1067–1078, Aug 2005.

[96] G. Contesso et al. The importance of histologic grade in long-term prognosis of breast cancer: a study of 1,010 patients, uniformly treated at the institut gustave-roussy. *J Clin Oncol*, 5(9):1378–1386, Sep 1987.

[97] R. F. Brem et al. Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. *AJR Am J Roentgenol*, 181(3):687–693, Sep 2003.

[98] S. Petushi et al. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. *BMC Med Img*, 6:14, 2006.

[99] A. Madabhushi et al. Automated detection of prostatic adenocarcinoma from high-resolution ex vivo mri. *IEEE Trans Med Imaging*, 24(12):1611–1625, Dec 2005.

[100] J. Sudbø, 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*, 21(2):71–86, 2000.

Vita

Scott Doyle

2011	Ph.D. in Biomedical Engineering, Rutgers University
2002-06	B.S. in Biomedical Engineering, Rutgers University
2002	Graduated from High Technology High School
2008-2011	Department of Defense Prostate Cancer Research Program Pre-doctoral Fellow, Department of Biomedical Engineering, Rutgers University
2006-2008	Research Assistant, Department of Biomedical Engineering, Rutgers University