A Boosted Bayesian Multiresolution Classifier for Prostate Cancer Detection From Digitized Needle Biopsies

Scott Doyle*, Michael Feldman, John Tomaszewski, and Anant Madabhushi, *Senior Member, IEEE*

***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 multiresolution (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 comprising 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, ten 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, respectively, at the lowest, intermediate, and highest resolution levels and 2) an eightfold savings in terms of computational time compared to running the algorithm directly at full (highest) resolution. The BBMR model outperformed (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.**

***Index Terms*—Computer-aided detection (CAD), histology, prostate cancer (CaP), quantification, supervised classification.**

I. Introduction

THE AMERICAN Cancer Society predicts that over 192 000 new cases of prostate cancer (CaP) will be diagnosed in the U.S. 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 [1]. Over one million prostate biopsies are performed annually in the U.S., 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 [2], 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 [3]. 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) [4], [5]. These pitfalls can be mitigated by introducing a quantitative "second reader" capable of automatically, accurately, and reproducibly finding suspicious CaP regions on the image [6]. 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 study 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 [7]. Computer-aided diagnosis (CAD) algorithms have been proposed for detecting neuroblastoma [8], identifying and quantifying extent of lymphocytic infiltration on breast biopsy tissue [9], and grading astrocytomas in brain biopsies [10]. In the context of detecting CaP on histopathology, earlier CAD approaches have employed low-level image characteristics, such as color, texture, and wavelets [11], second-order statistical [12], and morphometric attributes [13] in conjunction with classifier systems to distinguish benign from CaP regions. Diamond *et al*. [14] devised a system for distinguishing between stroma, benign epithelium, and CaP images measuring 100×100 pixels in size taken from whole-mount histology specimens. Using

Manuscript received December 3, 2009; revised March 17, 2010 and April 29, 2010; accepted May 3, 2010. Date of publication June 21, 2010; date of current version April 20, 2012. This work was supported by the Wallace H. Coulter Foundation, New Jersey Commission on Cancer Research, National Cancer Institute under Grant R01CA136535-01, Grant ARRA-NCI-3 R21 CA127186-02S1, Grant R21CA127186-01, and Grant R03CA128081-01, by the Department of Defense under Grant W81XWH-08-1-0145, by the Cancer Institute of New Jersey, Bioimagene, Inc., and by the Life Science Commercialization Award from Rutgers University. *Asterisk indicates corresponding author.*

*S. Doyle is with the Department of Biomedical Engineering, Rutgers University, Piscataway, NJ 08854 USA (e-mail: scottdo@eden.rutgers.edu).

M. Feldman and J. Tomaszewski are with the Department of Surgical Pathology, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: Michael.Feldman2@uphs.upenn.edu; John.Tomaszewski@uphs.upenn.edu).

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

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

Digital Object Identifier 10.1109/TBME.2010.2053540

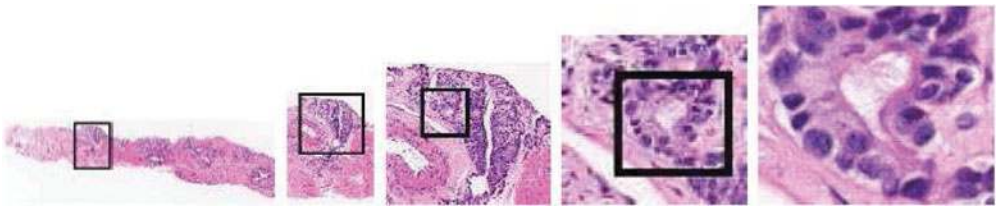

Fig. 1. Illustration of the multiresolution approach, where lower resolutions are used to identify suspicious regions that are later analyzed at higher resolution. This multiresolution approach results in significant computational savings. The most discriminatory features for CaP detection are learned and used to train a classifier at each image resolution.

morphological and texture features, an overall accuracy of 79.3% was obtained on 8789 samples, each of which represented a homogeneous section of tissue. Tabesh *et al*. [13] 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*. [15] 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 were obtained. Again, these results are achieved on preselected 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 [15]–[17]. The reliance on preselected 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 preextracted 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 multiresolution (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. Fig. 1 illustrates the scheme employed in this study 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 predefined 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 multiresolution image features employed by Viola and Jones [18], 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 [19], and wavelet features [20], [21]. 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 and 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 [21] will be more important (see Fig. 1).

Kong *et al*. [8] employed a similar multiresolution 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*. [23] used a multiclassifier 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 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*. [24] 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 outperform others. Madabhushi *et al*. [25] evaluated 14 different classifier ensemble schemes for the purpose of detecting CaP 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 classification performance, given identical training and testing data.

In our study, we have sought to select and extract features in a way that reflects 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 [22]. Each feature's weight is determined by how well the feature can discriminate between cancer and noncancer 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 multiresolution 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*. [26], who employed a similar AdaBoost technique to the classification of lung bronchovascular anatomy in computed tomography. 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) [27] or project the data into higher dimensional space (SVMs) [28], 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 noncancer classes.

Our methodology, called the boosted 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 10 000–50 000 pixels along each dimension) taken from 58 patient studies.

The rest of this paper is organized as follows. In Section II, we discuss our dataset and the initial preprocessing steps. In Section III, we discuss the feature extraction procedure. In Section IV, we describe the BBMR algorithm. Experimental design is described in Section V, and the results of analysis are presented in Section VI. Discussion of the results and concluding remarks are presented in Sections VII and VIII, respectively.

II. Brief Overview of Methodology and Preprocessing of Data

A. Image Digitization and Decomposition

An overview of our methodology is illustrated in Fig. 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 whole-slide scanning system operating at $40\times$ optical magnification. Images are saved to disk using the ImageScope software package as 8-bit tagged image file format files (scanner and software both from Aperio, Vista, CA). Tissue staining, fixing, and scanning were done at the Department of Surgical Pathology, University of Pennsylvania. The images digitized at the $40\times$ 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 GB.

An image pyramid was created using the pyramidal decomposition algorithm described by Burt and Adelson [29]. In this procedure, Gaussian smoothing is performed on the full-resolution $(40\times)$ image followed by subsampling of the smoothed image by a factor of 2. 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 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 I.

B. 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 color space captured by the scanner to the HSI 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 [30]. Thus, differences

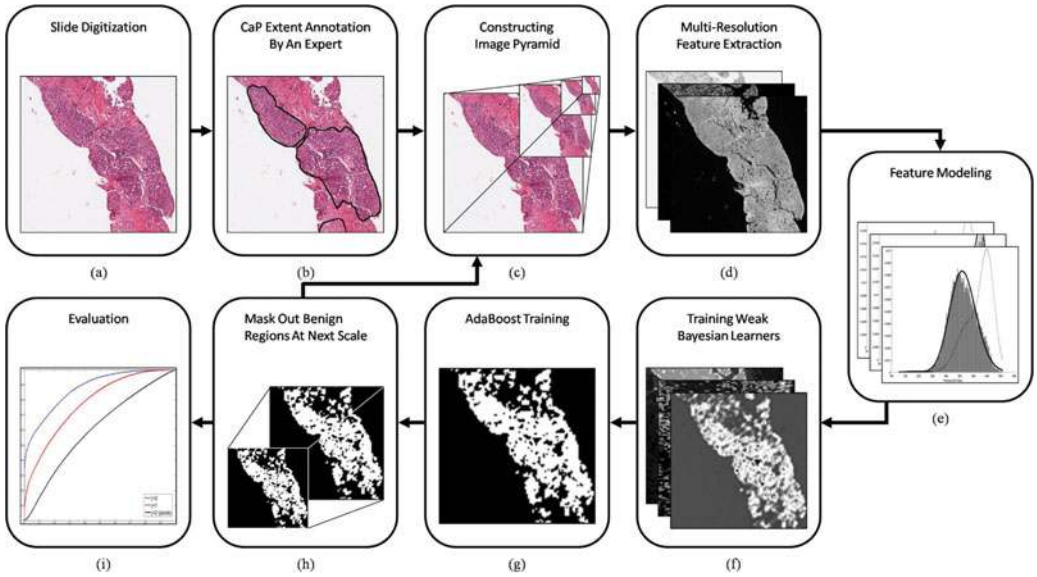

Fig. 2. Flowchart illustration of the working of the BBMR algorithm. (a) Slide digitization captures tissue samples at high resolution and (b) ground-truth regions of cancer are manually labeled. (c) Pyramidal decomposition is performed to obtain a set of successively smaller resolution levels. (d) At each level, several image features are extracted and (e) modeled via a Bayesian framework. (f) Weak classifiers thus constructed are combined using (g) the AdaBoost algorithm [22] into a single strong classifier for a specific resolution. (h) Probabilistic output of the AdaBoost classifier [22] 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.

TABLE I
DESCRIPTION OF THE DATASET, IMAGE PARAMETERS, GROUND-TRUTH ANNOTATION, AND PERFORMANCE MEASURES USED IN THIS STUDY

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

that naturally occur between different biopsy slides will be constrained to one channel instead of affecting all three.

C. 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 slide-viewing software. Labels were placed on the original scanned image and were propagated through the pyramid using the decomposition procedure described in Section II-A. 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 noncancer class, comprising benign epithelium and stroma, was also labeled for training; for evaluation, all noncancer regions (whether labeled as benign or unlabeled) were considered to be benign. Regions where both cancer and noncancerous 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 [see Fig. 3(d)].

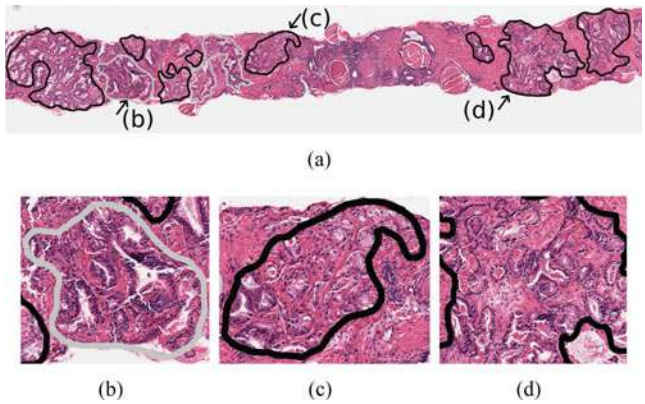

Fig. 3. (a) Original image with cancer (black contours) and noncancer (gray contour) regions labeled by an expert. (b) Closeup of the noncancer region. (c) and (d) Closeups of cancerous regions. Regions shown in (b), (c), and (d) are indicated on (a) by black arrows and text labels.

D. Notation

The notation used in this paper is summarized in Table II. We represent a digitized image by a pair $\mathcal{C} = (C, f)$, where C

TABLE II
LIST OF FREQUENTLY APPEARING NOTATION AND SYMBOLS IN THIS PAPER

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 Feature u
$P(\omega_i \mid f_u(c))$	*A posteriori* probability of class ω_i	$p(f_u(c) \mid \omega_i)$	Probability density function
O_c	Co-occurrence matrix for pixel c	$N_w(c)$	Window of size w around pixel c
κ, θ	Gabor filter parameters	$\mathbf{G}$	Gabor filter function
Γ	Gamma function	τ, η	Gamma function parameters
$\Pi^{\text{Ada}}(c)$	Classifier decision; $\{0, 1\}$	$\Pi_u(c)$	Weak classifier for pixel $c \in C$
T	Number of weak classifiers	δ_u	Weak 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 weight distribution
$\mathcal{A}(c)$	Ensemble result for pixel $c \in C$	$\mathcal{B}^j = (C^j, \Pi^{\text{Ada}})$	Binary classification scene

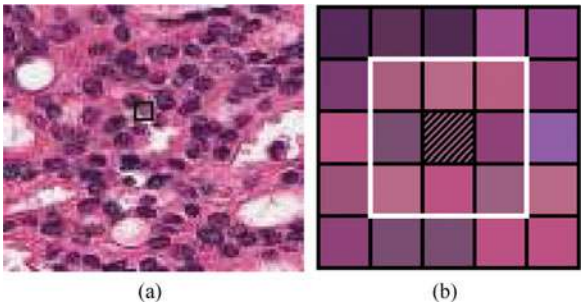

Fig. 4. Illustration of the procedure for calculating image features. (a) Magnified region of the original tissue image. (b) Pixelwise magnification of the region with the window N_w ($w = 3$) indicated by a white border and center pixel shaded with diagonal stripes.

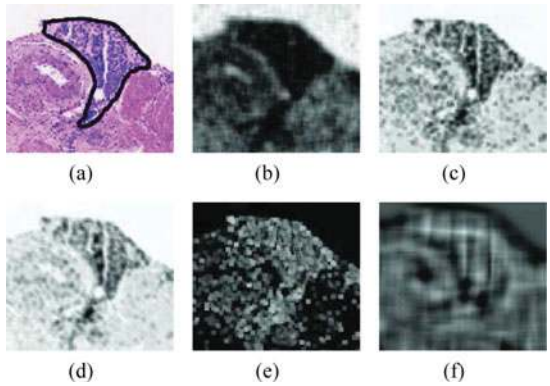

Fig. 5. (a) Original digitized prostate histopathological image with the manual segmentation of cancer overlaid (black contour), and five corresponding feature scenes. (b) Correlation ($w = 7$). (c) Sum variance ($w = 3$). (d) Gabor filter ($\theta = 5\pi/8$, $\kappa = 2$, $w = 3$). (e) Difference ($w = 3$). (f) Standard deviation ($w = 7$).

is a 2-D 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 $\mathcal{P} = \{\mathcal{C}^0, \mathcal{C}^1, \ldots, \mathcal{C}^{n-1}\}$, where $\mathcal{C}^j = (C^j, f)$ corresponds to the image at the jth level of the pyramid, where $j \in \{0, 1, \ldots, n-1\}$. We define the lowest (i.e., "coarsest") resolution level as $\mathcal{C}^0$ and the highest resolution level (at which 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 $\mathbf{F}(c) = [f_u(c) | u \in \{1, 2, \ldots, K\}]$, where $f_u(c)$ is the value of feature u at pixel $c \in C$. We denote as Φ_u, where $u \in \{1, 2, \ldots, K\}$, the random variable associated with each of the K features. An observation of Φ_u is made by calculating $f_u(c)$, for $c \in C$.

III. FEATURE EXTRACTION

The operations described in the following 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 Fig. 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 that 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 Fig. 5. The black contour in Fig. 5(a) represents the cancer region. Table III summarizes the image features extracted; details regarding the computation of the individual feature classes are given in the following.

1) *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 two diagonal axes, three 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.
2) *Co-occurrence Features:* Co-occurrence features [19] 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 $a, b \in \{1, \ldots, I\}$ 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 [19]. From O_c, a set of Haralick features (joint entropy, energy, inertia, inverse difference moment,

TABLE III
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

Feature Class and Individual Attributes	Parameters	Total Features
First-order Statistics (Average, Median, Standard Deviation, Range, Sobel, Kirsch, Gradient, Derivative)	Window size: $w \in \{3, 5, 7\}$	135
Co-occurrence Features (Joint Entropy, Energy, Inertia, Inverse Difference Moment, Correlation, Measurements of Correlation, Sum Average, Sum Variance, Sum Entropy, Difference Average, Difference Variance, Difference Entropy, Shade, Prominence, Variance)	Window size: $w \in \{3, 5, 7\}$ Distance: $\Delta = 1$	144
Gabor Features	Window size: $w \in \{3, 5, 7\}$ Frequency shift: $\kappa \in \{0, 1, \cdots, 7\}$ Orientation: $\theta \in \{\frac{\pi}{8}, \frac{2\pi}{8}, \cdots, \frac{8\pi}{8}\}$	648

Also included (right column) are the total number of features calculated for each feature class.

correlation, two measurements of correlation, sum average, sum variance, sum entropy, difference average, difference variance, difference entropy, dhade, 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.

3) *Steerable Filters:* The Gabor filter is constructed as a Gaussian function modulated by a sinusoid [21], [31]. 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

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

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 $\kappa \in \{0, 1, \ldots, 7\}$ and nine orientation parameter values ($\theta = \epsilon\pi/8$ where $\epsilon \in \{0, 1, \ldots, 8\}$), 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.

IV. BOOSTED BBMR CLASSIFIER

A. Bayesian Modeling of Feature Values

For each image feature-extracted (see Section III), 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 $u \in \{1, 2, \ldots, K\}$, $i \in \{1, 0\}$. We refer to the cancer class as ω_1 and the noncancer 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 [32]. 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 $u \in \{1, 2, \ldots, K\}$, we are interested in modeling the *a posteriori probability*, denoted by $P(\omega_i|\Phi_u)$, that feature values in Φ_u reflect class ω_i. This probability is given by the Bayes Rule [32]

$$P(\omega_i|\Phi_u) = \frac{P(\omega_i)p(\Phi_u|\omega_i)}{\sum_{k=0}^{1} P(\omega_k)p(\Phi_u|\omega_k)} \qquad (2)$$

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 as follows:

$$p(\Phi_u|\omega_i) \approx \Phi_u^{\tau-1} \frac{e^{-\Phi_u/\eta}}{\eta^\tau \Gamma(\tau)} \qquad (3)$$

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 Fig. 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 (3) models the cancer class distribution very accurately, some discrepancy between the model fit and the data for the noncancer class is observable in Fig. 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 noncancer class. Because all tissue data not labeled as cancer is considered part of the noncancer class, the noncancer class includes a diverse array of tissue types including stroma, normal epithelium, low- and high-grade prostatic intraepithelial neoplasia, atrophy, and inflammation [6], [33]. These diverse tissue types cause a high degree of variability in the noncancer 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 nontrivial task, limited by the time and expense required to obtain detailed annotations of these tissue classes.

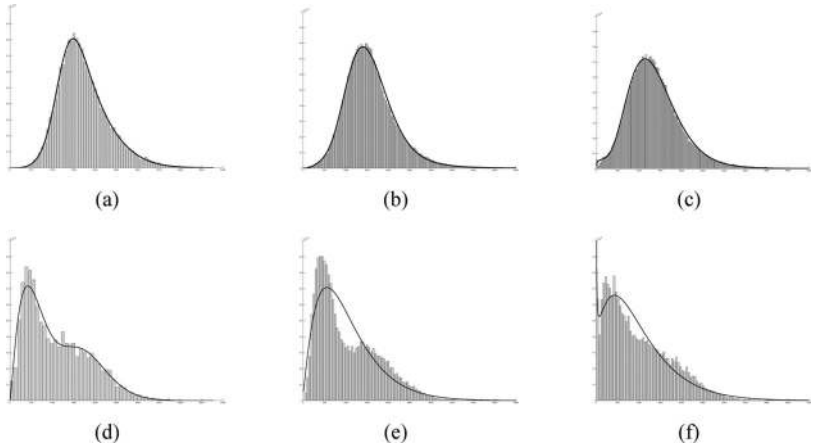

Fig. 6. PDFs for the Haralick variance feature for $w = 7$. Shown are the PDFs for resolutions levels (a) and (d) $j = 0$, (b) and (e) $j = 1$, and (c) and (f) $j = 2$. All PDFs in the top row [(a), (b), and (c)] are calculated for the cancer class, and in the bottom row [(d), (e), and (f)] for the noncancer 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.

B. Boosting Weak Classifiers

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

$$\Pi_u(c) = \begin{cases} 1, & \text{if } P(\omega_1 | f_u(c)) > \delta_u \\ 0, & \text{if } P(\omega_1 | f_u(c)) < \delta_u \end{cases} \tag{4}$$

where $\Pi_u(c) = 1$ corresponds to a positive (cancer) classification, $\Pi_u(c) = 0$ corresponds to a negative (noncancer) classification, and $\delta_u \in [0, 1]$ is a feature-specific threshold value. The optimal threshold value was learned offline on a set of training images using Otsu's thresholding method [34], a rapid method for choosing the optimal threshold by minimizing intraclass variance.

Once the weak classifiers, Π_u, for $u \in \{1, 2, \ldots, K\}$, have been constructed, they are combined to create a single strong classifier via the AdaBoost ensemble method [22]. 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 pixelwise classifiers $h_1, h_2, \ldots, h_T$, where $h_1(c) \in \{1, 0\}$ indicates the output of the highest weighted classifier and 2) T associated weights $\alpha_1, \alpha_2, \ldots, \alpha_T$ for each classifier. Note that $\alpha_1, \alpha_2, \ldots, \alpha_T$ 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

$$\mathcal{A}^{\mathrm{Ada}}(c) = \sum_{t=1}^{T} \alpha_t h_t(c). \tag{5}$$

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

$$\Pi^{\mathrm{Ada}}(c) = \begin{cases} 1, & \text{if } \mathcal{A}^{\mathrm{Ada}}(c) > \delta_{\mathrm{Ada}} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where δ_{Ada} is chosen using Otsu's method. For additional details on the AdaBoost algorithm (see [22]).

C. Multiresolution Implementation

The multiresolution 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 $\mathcal{B}^j = (C^j, \Pi^{\mathrm{Ada}})$, 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 multiresolution 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 study, we are not addressing the grading problem.

V. Experiments and Evaluation Methods

A. 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: 1) qualitative

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

TABLE IV
LIST OF THE DIFFERENT CLASSIFIERS COMPARED IN THIS STUDY

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

BBMR algorithm is denoted as ΠBBMR, while the additional classifiers are denoted with their feature estimation method (parametric (gamma) distribution or nonparametric feature distribution), as well as the different ensemble methods (AdaBoost, random forests, or none (single feature).

likelihood scene analysis; 2) area under the receiver operating characteristic (ROC) curve; and 3) classification accuracy at the pixel level (see Section V-C). Additional experiments were performed to explore different aspects of the BBMR algorithm. The list of experiments is as follows.

1) *Experiment 1 (Evaluation of BBMR Classifier):* We evaluated the output of the BBMR algorithm using the metrics listed in Section V-C, which include both qualitative and quantitative performance measures.
2) *Experiment 2 (Classifier Comparison):* We compared the BBMR classifier, denoted as Π^{BBMR}, with five other classifiers (summarized in Table IV). Two aspects of the system were altered to obtain the additional classifiers: a) the method of constructing the feature PDFs was changed from a Gamma distribution estimate (Section IV) to a nonparametric PDF obtained directly from the feature histograms (see the gray bars in Fig. 6) and b) The method used for combining weak classifiers was changed from the AdaBoost method to a randomized forest ensemble method [27]. Additionally, we tested a nonensemble approach, where the single best performing feature was used for classification. Different combinations of the method for generating the PDFs (parametric and nonparametric) and ensembles (AdaBoost, random, forests) yield the five additional classifiers shown in Table IV. While many additional ensemble and classification methods exist (for example, extremely randomized trees [35] is a recent alternative to random forests, and the SVM [28] is a classifier that does not employ a PDF), 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 IV 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.
3) *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.

B. Classifier Training

1) BBMR Classifier: To ensure robustness of BBMR to training data, randomized threefold cross-validation was performed on both a per-image and a per-patient basis.

1) *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 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, respectively, 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 noncancer 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 noncancer 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 recreated using features calculated from pixels in the G_1 and G_3 groups. As before, equal numbers of cancer and noncancer samples are used to generate the training set, while all of the pixels in G_2 that have not been classified as noncancer at an earlier 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.
2) *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.

2) 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 V-C. A total of T trees were used in the ensemble, each of which used a maximum of K/T randomly selected features to generate the tree. For each of the trees, the C4.5 algorithm [36] was employed to construct and prune the tree. Each tree was optimally pruned to a length that helped maximize classifier accuracy.

C. 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 noncancer pixels, for all $c \in C$. We determine three methods for classifier evaluation: 1) likelihood scene analysis; 2) area under the ROC curve (AUC); and 3) accuracy.

1) Comparative Analysis of Classifier-Generated CaP Probability: We obtain a likelihood scene $\mathcal{L}^j = (C^j, \mathcal{A}^{\mathrm{Ada}})$ for image resolution level $j \in \{0, 1, 2\}$, where each pixel's value is given by $\mathcal{A}^{\mathrm{Ada}}(c)$ (5). Likelihood scenes are compared with pathologist-defined ground truth for a qualitative evaluation of classifier performance.

2) Area Under the ROC Curve (AUC): Classifier sensitivity (SENS) and specificity (SPEC) in detecting disease extent is determined by varying δ_{Ada} (see Section IV-B). For a specific threshold δ_{Ada}, and for all $c \in C$, the number of true positives ($\mathrm{TP}_{\delta_{\mathrm{Ada}}}$) is found as $|\{c \in C | g(c) = \Pi^{\mathrm{Ada}}(c) = 1\}|$, false positives ($\mathrm{FP}_{\delta_{\mathrm{Ada}}}$) is $|\{c \in C | g(c) = 0, \Pi^{\mathrm{Ada}}(c) = 1\}|$, true negatives ($\mathrm{TN}_{\delta_{\mathrm{Ada}}}$) is $|\{c \in C | g(c) = \Pi^{\mathrm{Ada}}(c) = 0\}|$, and false negatives ($\mathrm{FN}_{\delta_{\mathrm{Ada}}}$) is $|\{c \in C | g(c) = 1, \Pi^{\mathrm{Ada}}(c) = 0\}|$, 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. $\mathrm{SENS}_{\delta_{\mathrm{Ada}}}$ and $\mathrm{SPEC}_{\delta_{\mathrm{Ada}}}$ can then be determined as

$$\mathrm{SENS}_{\delta_{\mathrm{Ada}}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$\mathrm{SPEC}_{\delta_{\mathrm{Ada}}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{8}$$

By varying the threshold as $0 \leq \delta_{\mathrm{Ada}} \leq \max[\mathcal{A}^{\mathrm{Ada}}]$, ROC curves for all the classifiers considered can be plotted by varying sensitivity versus $1 -$ 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 noncancer classes.

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

$$\mathrm{ACC}_{\delta_{\mathrm{Ada}}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} = \frac{\mathrm{TP} + \mathrm{TN}}{|C|}. \tag{9}$$

For our evaluation, we choose δ_{Ada} as described in Section IV-B, 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 tradeoff 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).

VI. Experimental Results

A. Experiment 1: Evaluation of BBMR Classifier

Figs. 7 and 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 Figs. 7(b) and 8(b) are the magnified image areas corresponding to the cancer region (as determined by the pathologist), and shown in Figs. 7(c) and 8(c) are the noncancer image areas. As the image resolution increases, more benign regions are pruned and eliminated at the lower resolutions.

B. 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 IV is shown for the image-based cross-validation experiments in Table V. Shown in Table VI are the 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 threefold cross-validation. Fig. 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 Fig. 9, there is a qualitative comparison of the three different feature ensemble methods: Fig. 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 Figs. 9(b)–(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 propagated at the next level, decreasing overall accuracy.

C. Experiment 3: BBMR Parameter Analysis

1) AdaBoost Ensemble Size T: The graph in Fig. 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

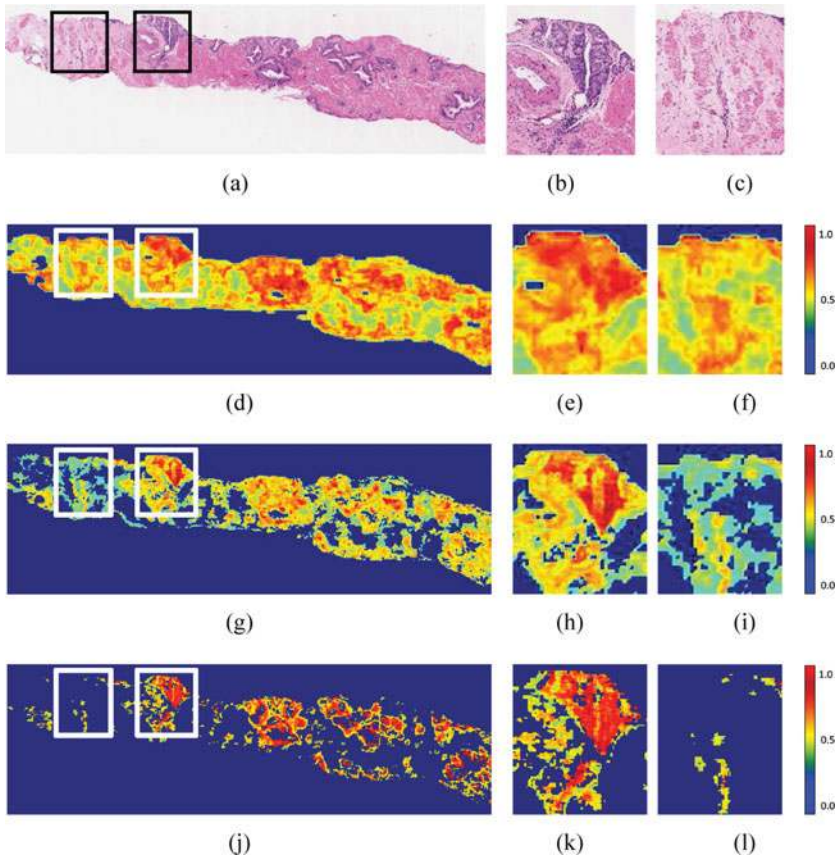

Fig. 7. Illustration of CaP classification via Π^{BBMR} on a single prostate image sample. The full image is shown in (a), with corresponding likelihood scenes $\mathcal{L}^0$, $\mathcal{L}^1$, and $\mathcal{L}^2$ shown in (d), (g), and (j), respectively. Closeups of cancer and benign regions (indicated 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.

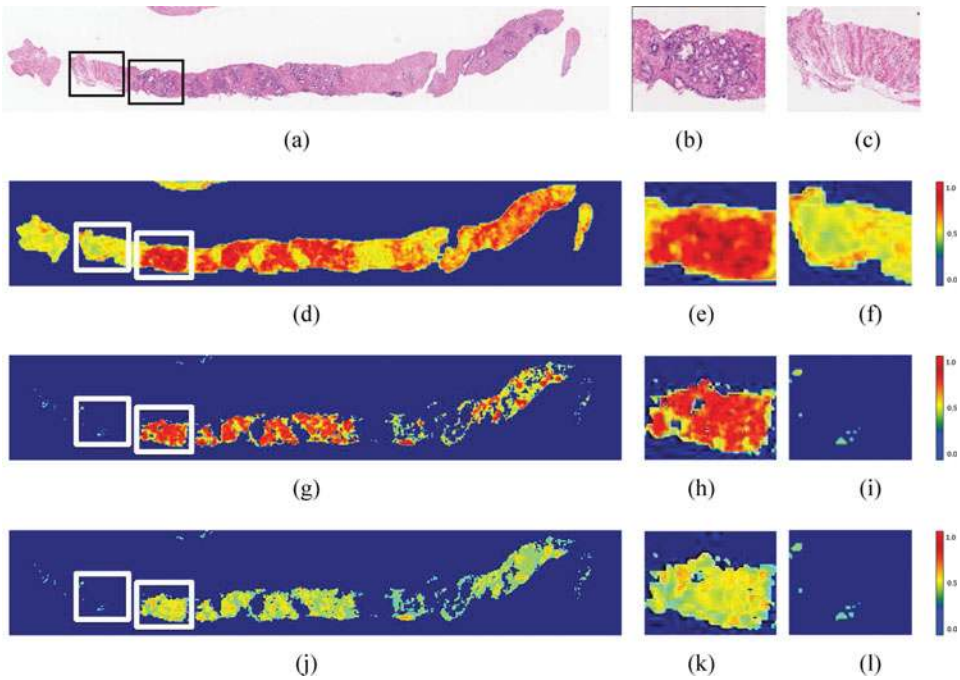

Fig. 8. Illustration of CaP classification via Π^{BBMR} on a single prostate image sample. The full image is shown in (a), with corresponding likelihood scenes $\mathcal{L}^0$, $\mathcal{L}^1$, and $\mathcal{L}^2$ shown in (d), (g), and (j), respectively. Closeups of cancer and benign regions are shown in (b) and (c), with corresponding CaP classification shown in subsequent rows as for the full image.

TABLE V
IMAGE-BASED CROSS-VALIDATION RESULTS

	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)
$\Pi^{\mathrm{RF,gamma}}$	0.33 (0.03)	0.46 (0.01)	0.37 (0.11)	0.50 (0.01)	0.56 (0.04)	0.31 (0.04)
$\Pi^{\mathrm{best,gamma}}$	0.54 (0.06)	0.56 (0.10)	0.54 (0.05)	0.55 (0.11)	0.55 (0.04)	0.55 (0.14)
$\Pi^{\mathrm{Ada,feat}}$	0.32 (0.04)	0.53 (0.05)	0.25 (0.11)	0.53 (0.05)	0.20 (0.04)	**0.69 (0.16)**
$\Pi^{\mathrm{RF,feat}}$	0.73 (0.01)	0.50 (0.02)	0.66 (0.01)	0.52 (0.03)	0.63 (0.01)	0.60 (0.01)
$\Pi^{\mathrm{best,feat}}$	0.44 (0.03)	0.52 (0.01)	0.26 (0.01)	0.46 (0.03)	0.43 (0.01)	0.54 (0.01)

Shown are ACC and AUC values for all the classifiers considered in this study (listed in Table IV) at each of the three 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.

TABLE VI
PATIENT-BASED CROSS-VALIDATION RESULTS

	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)**
$\Pi^{\mathrm{RF,feat}}$	0.79 (0.03)	0.63 (0.07)	0.73 (0.01)	0.55 (0.01)	0.59 (0.01)	0.51 (0.03)

Shown are ACC and AUC values for Π^{BBMR} and $\Pi^{\mathrm{RF,feat}}$ at each of the three 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.

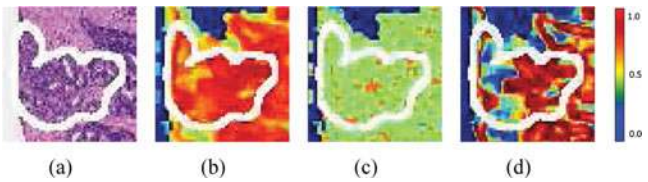

Fig. 9. Qualitative comparison of classifier performance on a single prostate image sample. The original image is shown in (a) with the cancer region denoted in a white contour. Likelihood scenes corresponding to Π^{BBMR}, $\Pi^{\mathrm{RF,gamma}}$, and $\Pi^{\mathrm{best,gamma}}$ are shown in (b)–(d), 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.

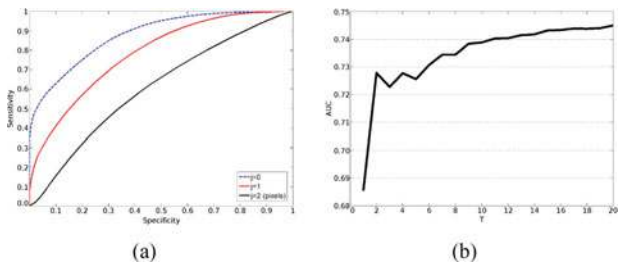

Fig. 10. (a) ROC curves generated at $j = 0$ (dashed blue line), $j = 1$ (solid red line), and $j = 2$ (solid black line) using the BBMR classifier. The apparent decrease in classifier (BBMR) accuracy with increasing image resolution is due to a lack of granularity in image annotation at the higher resolutions (see Fig. 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 $j = 1$.

classifier ensemble. The independent axis in Fig. 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 Fig. 10(b), T was varied from 1 to 20, and μ_{AUC} remained relatively stable beyond $T = 10$. The trends shown in Fig. 10(b) for $j = 2$ were also observed for $j = 0$ and $j = 1$.

2) AdaBoost Feature Selection: The features chosen by AdaBoost at each resolution level are specific to the information available at that image resolution. Table VII shows the top five features selected by AdaBoost at each image resolution. Table VII 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 gray-level 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 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.

3) Computational Efficiency: Fig. 11 illustrates the computational savings in employing the multiresolution BBMR scheme. The nonmultiresolution-based approach employs approximately four 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 study and for images of approximately 1000×1000 pixels, the analysis of a single image requires less than 3 min on average. All computations

TABLE VII
LIST OF THE TOP FIVE FEATURES CHOSEN BY ADABOOST AT THE THREE RESOLUTION LEVELS

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

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.

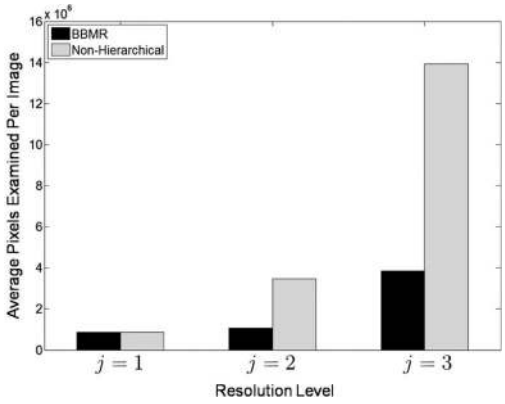

Fig. 11. Efficiency of the system using the BBMR system (black) and a nonhierarchical method (gray), measured in terms of the number of pixel-level calculations for levels $j=0$, $j=1$, and $j=2$.

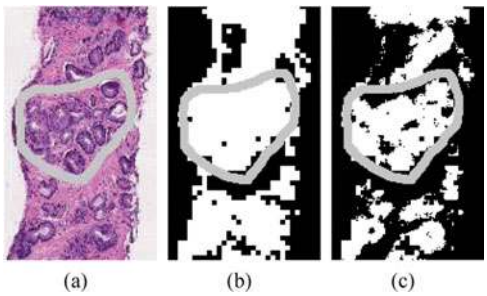

Fig. 12. Qualitative results illustrating the apparent dropping off in pixel-level classifier accuracy at higher image resolutions. (a) Original image with the cancer region annotated in gray. Binary image in (b) shows the overlay of the BBMR classifier detection results for CaP at $j=0$ and (c) shows the 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.

in this study were done on an dual core Xeon 5140 2.33-GHz computer with 32-GB RAM running the Ubuntu Linux operating system and MATLAB software package version 7.7.0 (R2008b) (The MathWorks, Natick, Massachusetts).

VII. DISCUSSION

Examining the ROC curves in Fig. 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

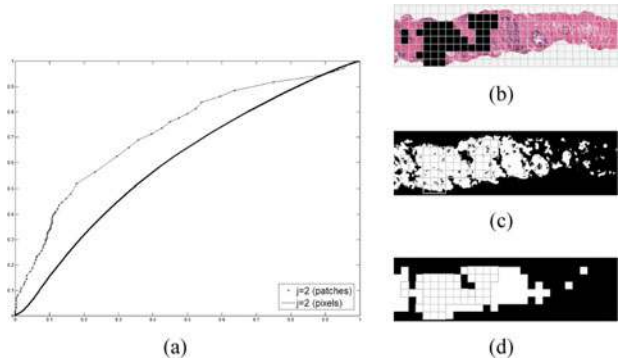

Fig. 13. (a) Comparison of ROC curves between pixel-based classification (solid black line) and patch-based classification (black dotted line). (b) Original image with a uniform 30-by-30 grid superimposed. Black boxes indicate the cancer region. (c) Pixel-wise classification results at resolution level $j=2$, 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.

resolutions $j=0$ and $j=2$ (see Fig. 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 $j=2$ (see Fig. 12(b) and (c), respectively).

As is discernible in Fig. 12(a), the cancer region annotated by the expert is heterogeneous, comprising many benign structures (stroma, intragland 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 [see Fig. 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 patchwise classification on a sample image at level $j = 2$ are shown in Fig. 13(d) and compared with the BBMR pixel-level classifier on the same image in Fig. 13(c). Intertwined regions of benign tissue within the diseased areas, classified as benign by the pixelwise BBMR classifier (and labeled as "false negatives" as a result) are classified as cancerous by the patchwise BBMR classifier. This yields an ROC curve with a greater area; compare corresponding curves for the pixel-based and patch-based BBMR classifiers in Fig. 13(a).

We would like to point out that the apparent drop-off in classifier accuracy has to do with the lack of ground-truth granularity at higher resolutions. Pathologists are able to distinguish cancer from noncancer 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 study. In a Gleason grading system [13], [37], such additional high-level features will be calculated from the suspicious regions detected at the end of the BBMR classification algorithm.

VIII. Concluding Remarks

In this study, we presented a boosted BBMR classifier for automated detection of CaP from digitized histopathology, a necessary precursor to automated Gleason grading. To the best of our knowledge, this study 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 multiresolution 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 multiresolution framework reduces the amount of time needed to analyze large images by approximately 4–6 times compared to a nonmultiresolution-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 nonparametric formulation for feature modeling.

Pixelwise 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.

References

[1] B. Matlaga, L. Eskew, and D. McCullough, "Prostate biopsy: Indications and technique," *J. Urol*., vol. 169, no. 1, pp. 12–19, 2003.

[2] H. Welch, E. Fisher, D. Gottlieb, and M. Barry, "Detection of prostate cancer via biopsy in the medicare-seer population during the psa era," *J. Nat. Cancer Inst*., vol. 99, no. 18, pp. 1395–1400, 2007.

[3] D. Gleason, "Classification of prostatic carcinomas," *Cancer Chemother. Rep*., vol. 50, no. 3, pp. 125–128, 1966.

[4] D. Bostwick and I. Meiers, "Prostate biopsy and optimization of cancer yield," *Eur. Urol*., vol. 49, no. 3, pp. 415–417, 2006.

[5] W. C. Allsbrook, K. A. Mangold, M. H. Johnson, R. B. Lane, C. G. Lane, and J. I. Epstein. (2001). Interobserver reproducibility of gleason grading of prostatic carcinoma: General pathologist. *Human Pathol*. [Online]. 32(1), pp. 81–88. Available: http://dx.doi.org/10.1053/hupa.2001.21135

[6] J. Epstein, P. Walsh, and F. Sanfilippo, "Clinical and cost impact of second-opinion pathology. review of prostate biopsies prior to radical prostatectomy," *Amer. J. Surg. Pathol*., vol. 20, no. 7, pp. 851–857, 1996.

[7] 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," *Exp. Biol. Med*., vol. 234, pp. 860–879, 2009.

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

[9] A. Basavanhally, S. Ganesan, S. Agner, J. Monaco, G. Bhanot, and A. Madabhushi, "Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology," *IEEE Trans. Biomed. Eng*., vol. 57, no. 3, pp. 642–653, Mar. 2010.

[10] 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," *Comput. Methods Programs Biomed*., vol. 90, no. 3, pp. 251–261, 2008.

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

[12] A. Esgiar, R. Naguib, B. Sharif, M. Bennett, and A. Murray, "Fractal analysis in the detection of colonic cancer images," *IEEE Trans. Inf. Technol. Biomed*., vol. 6, no. 1, pp. 54–58, Mar. 2002.

[13] A. Tabesh, M. Teverovskiy, H. Pang, V. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and gleason grading of histological images," *IEEE Trans. Med. Imag*., vol. 26, no. 10, pp. 1366–1378, Oct. 2007.

[14] J. Diamond, N. Anderson, P. Bartels, R. Montironi, and P. Hamilton, "The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia," *Human Pathol*., vol. 35, no. 9, pp. 1121–1131, 2004.

[15] R. Farjam, H. Soltanian-Zadeh, K. Jafari-Khouzani, and R. Zoroofi, "An image analysis approach for automatic malignancy determination of prostate pathological images," *Cytometry Part B (Clin. Cytometry)*, vol. 72, no. B, pp. 227–240, 2007.

[16] P. Huang and C. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE Trans. Med. Imag.*, vol. 28, no. 7, pp. 1037–1050, Jul. 2009.
[17] K. Jafari-Khouzani and H. Soltanian-Zadeh, "Multiwavelet grading of pathological images of prostate," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 6, pp. 697–704, Jun. 2003.
[18] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[19] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
[20] D. Gabor, "Theory of communication," *Proc. Inst. Electr. Eng.*, vol. 93, no. 26, pp. 429–457, 1946.
[21] B. Manjunath and W. Ma, "Texture features for browsing and retrieval of image data," *Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, 1996.
[22] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Mach. Learning*, 1996, pp. 148–156.
[23] A. Sboner, C. Eccher, E. Blanzieri, P. Bauer, M. Cristofolini, G. Zumiani, and S. Forti, "A multiple classifier system for early melanoma diagnosis," *Artif. Intell. Med.*, vol. 27, pp. 29–44, 2003.
[24] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M. Chung, and S. Johnson, "Spatially augmented lpboosting for ad classification with evaluations on the adni dataset," *Neuroimage*, vol. 48, pp. 138–149, 2009.
[25] A. Madabhushi, J. Shi, M. Feldman, M. Rosen, and J. Tomaszewski, "Comparing ensembles of learners: Detecting prostate cancer from high resolution MRI," *Comput. Vis. Methods Med. Image Anal.*, vol. LNCS 4241, pp. 25–36, 2006.
[26] R. Ochs, J. Goldin, F. Abtin, H. Kim, K. Brown, P. Batra, D. Roback, M. McNitt-Gray, and M. Brown, "Automated classification of lung bronchovascular anatomy in ct using adaboost," *Med. Image Anal.*, vol. 11, pp. 315–324, 2007.
[27] L. Breiman, "Random forests," *Mach. Learning*, vol. 45, pp. 5–32, 2001.
[28] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learning*, vol. 20, pp. 273–297, 1995.
[29] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. COM-31, no. 4, pp. 532–540, Apr. 1983.
[30] H. Cheng, X. Jiang, Y. Sun, and J. Wang, "Color image segmentation: Advances and prospects," *Pattern Recognit.*, vol. 34, no. 12, pp. 2259–2281, 2001.
[31] A. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 1990, pp. 14–19.
[32] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York: Wiley, 2001.
[33] N. Borley and M. Feneley, "Prostate cancer: Diagnosis and staging," *Asian J. Androl.*, vol. 11, pp. 74–80, 2009.
[34] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
[35] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learning*, vol. 63, pp. 3–42, 2006.
[36] J. Quinlan, "Decision trees and decision-making," *IEEE Trans. Syst., Man Cybern.*, vol. 20, no. 2, pp. 339–346, Mar./Apr. 1990.
[37] S. Doyle, S. Hwang, K. Shah, J. Tomaszewski, M. Feldman, and A. Madabhushi, "Automated grading of prostate cancer using architectural and textural image features," in *Proc. ISBI*, 2007, pp. 1284–1287.

Scott Doyle received his Ph.D. degree in biomedical engineering in 2011 from Rutgers University, Piscataway, NJ. His research focus is on developing computer decision-support systems for pathologists to quantitatively analyze patient data, detect and diagnose disease, and develop optimal treatment plans. Currently, he is employed as the Director of Research at Ibris, Inc., a start-up founded to commercialize the research performed during the course of his thesis work.

Dr. Doyle has been awarded the Department of Defense Predoctoral Fellowship award for his prostate cancer research.

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

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

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

He is currently the Chief of Pathology at the University of Buffalo. He is a nationally recognized expert in diagnostic genitourinary pathology. His current research interests include high-resolution MRI of prostate and breast cancer, computer-assisted diagnosis, and high-dimensionality data fusion in the creation of a new diagnostic testing paradigms.

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

He is currently an Associate Professor in the Department of Biomedical Engineering, Rutgers University, Piscataway, NJ, where he is also the Director of the Laboratory for Computational Imaging and Bioinformatics. He is also a member of the Cancer Institute of New Jersey and an Adjunct Assistant Professor of radiology at the Robert Wood Johnson Medical Center, New Brunswick, NJ. He has authored or coauthored more than 130 papers published in various international journals and conferences and book chapters, and has several patents pending in medical image analysis, computer-aided diagnosis, machine learning, and computer vision.

Dr. Madabhushi is the recipient of a number of awards for research as well as teaching, including the Coulter Phase 1 and Phase 2 Early Career Award (2006 and 2008), the Excellence in Teaching Award (2007–2009), the Society for Imaging Informatics in Medicine New Investigator Award (2008), and the Life Sciences Commercialization Award (2008). Recently he co-founded a company, Ibris, Inc., to commercialize work that has been performed in the lab.