# Visually Meaningful Histopathological Features for Automatic Grading of Prostate Cancer

M. Khalid Khan Niazi, Keluo Yao, Debra L. Zynger, Steven K. Clinton, James Chen, Mehmet Koyutürk, Thomas LaFramboise, and Metin Gurcan

***Abstract*—Histopathologic features, particularly Gleason grading system, have contributed significantly to the diagnosis, treatment, and prognosis of prostate cancer for decades. However, prostate cancer demonstrates enormous heterogeneity in biological behavior, thus establishing improved prognostic and predictive markers is particularly important to personalize therapy of men with clinically localized and newly diagnosed malignancy. Many automated grading systems have been developed for Gleason grading but acceptance in the medical community has been lacking due to poor interpretability. To overcome this problem, we developed a set of visually meaningful features to differentiate between low- and high-grade prostate cancer. The visually meaningful feature set consists of luminal and architectural features. For luminal features, we compute: 1) the shortest path from the nuclei to their closest luminal spaces; 2) ratio of the epithelial nuclei to the total number of nuclei. A nucleus is considered an epithelial nucleus if the shortest path between it and the luminal space does not contain any other nucleus; 3) average shortest distance of all nuclei to their closest luminal spaces. For architectural features, we compute directional changes in stroma and nuclei using directional filter banks. These features are utilized to create two subspaces; one for prostate images histopathologically assessed as low grade and the other for high grade. The grade associated with a subspace, which results in the minimum reconstruction error is considered as the prediction for the test image. For training, we utilized 43 regions of interest (ROI) images, which were extracted from 25 prostate whole slide images of The Cancer Genome Atlas (TCGA) database. For testing, we utilized an independent dataset of 88 ROIs extracted from 30 prostate whole slide images. The method resulted in 93.0% and 97.6% training and testing accuracies, respectively, for the spectrum of cases considered. The application of visually meaningful features provided promising levels of accuracy and consistency for grading prostate cancer.**

***Index Terms*—Color deconvolution, cytological and architectural features, geodesic distance, shortest path.**

## I. Introduction

ALTHOUGH prostate cancer is the second most common cancer in males in the United States, the risk of dying from prostate cancer is several-fold lower. An average male's lifetime risk of prostate cancer is approximately 11%, while the risk of death from prostate cancer risk is 4% [1]. The discrepancy could be due to our inability to distinguish those with indolent disease and those with a more aggressive variant, especially in the early stages. This lack of prognostic ability results in unnecessary and potentially harmful treatment. A widely used prognostic method relies on histopathologic information derived from the tumor identified in the prostate needle core biopsy known as the Gleason score. Developed by Dr. Donald Gleason in the 1960s, Gleason scoring relies on identifying changes in primarily glandular architecture of prostate and is applied manually using light microscopy on hematoxylin and eosin (H&E) stained glass tissue by pathologists [2]. In the absence of more reliable biomarkers, the evaluation based on a Gleason score from needle biopsy is utilized along with other clinicopathologic features such as serum prostate-specific antigen, often as components of nomograms, to risk stratify patients as an aid to define treatment options [3]. A key goal is to simultaneously reduce overtreatment of indolent cancer and its complications while providing appropriate curative therapy and options for experimental strategies in those at greatest risk. Previous studies have demonstrated significant interreader variability among practicing pathologists depending on experience and training [4]–[6]. Prognostication, treatment decisions, and research-oriented stratification made based on the reported Gleason score may be inappropriate [7]. However, recent advances seem to indicate significant prognostic differences even within intermediate grades such as Gleason score 7 [8], indicating a need for precise and accurate histologic evaluation. The objective in this study is to develop, calibrate, and validate visually meaningful features to provide automated criteria for prostate risk grading.

The literature contains a plethora of methodology to automate the Gleason grading system [9]. In [10], Diamond *et al.* presented an automated method to classify prostate cancer images into tumor, stroma, and normal regions based on Haralick features. To discriminate moderately from poorly differentiated

Manuscript received October 5, 2015; revised February 4, 2016 and March 29, 2016; accepted April 30, 2016. Date of publication May 17, 2017; date of current version June 29, 2017.

M. K. K. Niazi, J. Chen, and M. Gurcan are with the Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210 USA (e-mail: muhammad.niazi@osumc.edu; James.Chen@osumc.edu; Metin.Gurcan@osumc.edu).

K. Yao and D. L. Zynger are with the Department of Pathology, The Ohio State University, Columbus, OH 43210 USA (e-mail: Keluo.Yao@osumc.edu; Debra.Zynger@osumc.edu).

S. K. Clinton is with the Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210 USA (e-mail: Steven.Clinton@osumc.edu).

M. Koyutürk is with the Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH 44106 USA (e-mail: mxk331@case.edu).

T. LaFramboise is with the Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH 44106 USA (e-mail: Thomas.LaFramboise@case.edu).

Digital Object Identifier 10.1109/JBHI.2016.2565515

cancers, the authors extracted statistical and structural features of the nuclei [11]. These features were extracted by using a partly trained multilayer neural network. The classification was performed with the quadratic Bayesian classifier. The eigenvalue analysis of the Fourier transformed prostate images was utilized in [12] to replicate the Gleason grading. In [13], architectural features were derived by constructing spanning trees. These trees were generated by connecting the cell nuclei over a prostate image. The Gleason scoring was performed by matching the properties of these trees to the training trees in the database. In [14], energy and entropy features were extracted from multiwavelet coefficients of the prostate image. It was followed by simulated annealing to determine the most discriminating features. The resulting features were assigned a prostate score based on a *k*-nearest neighbor classifier.

In another effort [15] to automate the Gleason scoring, the glands were extracted using texture features and *k*-means clustering. Then the tree structure architecture (similar to classification trees) was applied along with a linear classifier to classify the image into Gleason scores. The tree structure architecture relied on the shape and intensity-based features of the glands. In [16], a Gleason grading system was presented which incorporated low-level knowledge, high-level knowledge, and structural constraints via domain knowledge to segment the prostate images. Morphological and architectural features from these segmented images were used to perform Gleason grading. A texton-based Gleason scoring was presented in [17]. The authors combined textons with support vector machines to classify pixels into Gleason patterns 3 and 4. In [18], Sparks and Madabhushi presented a statistical shape model to perform manifold regularization and its application to differentiate between Gleason patterns 3 and 4.

All of these methods were reported to have a reasonable accuracy. However, none of these methods presents visually meaningful and clinically interpretable features, especially to pathologists. Moreover, these methods do not add anything new to the pathologist's knowledge. Unfortunately, these methods lack the necessary ingredients to:

1) Allow the pathologist to interpret a certain score by visually meaningful features.
2) Enable the pathologists to share their findings with peers/residents/students in clinically relevant language.

In the current effort, we address these shortcomings.

The paper is organized as follows. The segmentation and feature extraction methodologies are presented in Section II. Section III presents the databases used in this study. The results are presented in Section IV followed by the discussion in Section V. Conclusions from this study are presented in Section VI.

## II. Method

### A. Segmentation

For segmentation of H&E-stained prostate tissue sample images, we have adopted a texture-based methodology to classify biologically meaningful components from the region of interest (ROI) images of the prostate tissue, inspired by the statistical approach of Varma and Zisserman [19]. Our objective is to automatically construct a deconvolution matrix [20] to perform segmentation of prostate images from a single channel. The method relies on texture instead of color to derive the color deconvolution matrix, and automatically establishes a correspondence between the hematoxylin-stained, eosin-stained, and background pixels in the training image to those in the test image. Once the correspondence is established, pixels in each group are individually averaged to compute the color deconvolution matrix.

In their work, Varma and Zisserman [19] modeled texture by joint probability distribution of filter responses. Each distribution was represented by the cluster centroids (textons). Their approach can be divided into three stages: 1) texton dictionary generation; 2) model generation; and 3) classification. During the first stage, the filter responses from the training images of the same texture class are vector quantized by modeling the probability density functions in the filter response space with the texton distributions. In the second (i.e., model generation) stage, the model of each training image (within the same class) is generated by first convolving it with the same set of filters in the first stage and then labeling each filter response with the texton that lies closest to it in the filter response space. The frequency, with which each texton occurs in the labeling, forms the model, i.e., a histogram of the occurrence of cluster centroids. This results in as many models as the number of training images in each texture class. In the third stage, classification, a test image is classified by forming its histogram and then choosing the closest model histogram learned from the training set.

At the first stage, one may choose from multiple filter banks [19], [21], [22]. Each of these filter banks are reported to have similar performances on several texture databases. For this study, we opted to use maximum response filter bank, MR8, which consists of 38 filters [23]. The MR8 filter bank contains filters at multiple orientations, but their filter responses are "collapsed" by recording only the maximum filter response across all orientations [19], resulting in rotational invariance. The filter bank consists of an isotropic Gaussian and an isotropic Laplacian of Gaussian filter, an edge filter at three scales and a bar filter at the same three scales. Both edge and bar filters are directional in nature and contain six orientations at each scale. Measuring only the maximum response across orientations reduces the number of responses from 38 to 8.

The segmentation method starts by applying an MR8 filter bank to the red channel of the H&E prostate images. The red channel of the RGB prostate image is selected as it provides relatively higher contrast between the nuclei and the stroma. The segmentation method uses a set of five training images. Each training image $T_x$ is manually segmented into nuclei ($N$), stroma ($S$), and luminal spaces ($L$) and the filtered response $F_{T_x}$ is clustered into $k$ clusters using *k*-means clustering. To automatically determine $k$, the number of clusters is sequentially increased from three to 15 to find a cluster $\mathrm{W}_N$ that maximizes $\#\{\mathrm{W}_N \cap^{N_s} 3$ and minimizes $\#\{(\mathrm{W}_N \cap^S) \cup (\mathrm{W}_N \cap^L)\}$, simultaneously, where $N_s$ is a subset of $N$, i.e., $N_s \subseteq N$. Here # represents the cardinality of a set, and $\cap$ the set intersection. Based on our experiments, the optimal values of $k$ varies between eight to 10 for the training images. Once $k$ is de-

termined, we pick another cluster $\mathrm{W}_S$ from the remaining $k-1$ clusters that maximizes $\#\{\mathrm{W}_S \cap^{S_s}\}$ and minimizes $\#\{(\mathrm{W}_S \cap^N) \cup (\mathrm{W}_S \cap^L)\}$, simultaneously. Similarly, we pick another cluster $\mathrm{W}_L$ from the remaining $k-2$ clusters that maximizes $\#\{\mathrm{W}_L \cap^{L_s}\}$ and minimizes $\#\{(\mathrm{W}_L \cap^N) \cup (\mathrm{W}_L \cap^S)\}$, simultaneously where, $S_S \subseteq S$ and $L_S \subseteq L$. It is obvious from the discussion that the selection of $k$ is mainly driven by $\mathrm{W}_N$. This points toward the directional nature of MR8 filter bank, which require a higher value of $k$ to separate circular objects (nuclei) than the linear objects (stroma), i.e., it is relatively more difficult for MR8 filter bank to cluster circular objects than linear objects. For this reason, we maximize $\#\{\mathrm{W}_S \cap^{N_S}\}$ instead of $\#\{\mathrm{W}_S \cap^N\}$ as it is difficult to cluster the whole nuclei into a single cluster. Let $C_x = \{C_{x1}, C_{x2}, \ldots C_{\mathrm{xk}}\}$ represent the centroids of $T_x$ where $x \in \{1,2,3,4,5\}$. $C_x$ is a set of $k$ eight-dimensional feature vectors. For any test image $V_y$, we compute the filtered response $F_{V_y}$, and compute the Euclidean distance, $Dis$, between each pair in $F_{V_y}$ and $C_x$. For each point in $F_{V_y}$, $Dis$ helps to determine the closest centroid in $C_x$, i.e., to establish correspondence between points (pixels) in $F_{V_y}$ and cluster centroids in $C_x$. This whole process is graphically illustrated in Fig. 1.

Once the correspondence is established, we create $\boldsymbol{F}_{\boldsymbol{V}_{\mathrm{yx}}}$ by replacing each point in $\boldsymbol{F}_{\boldsymbol{V}_y}$ with its corresponding centroid from $\boldsymbol{C}_x$. Now, the $\boldsymbol{C}_x$, which results in minimum reconstruction error $\boldsymbol{E}$ is chosen as the representative centroid for $\boldsymbol{F}_{\boldsymbol{V}_y}$

$$\boldsymbol{E} = \mathbf{min}_{\mathbf{x}} \left( \sum \left( \boldsymbol{F}_{\boldsymbol{V}_{\mathrm{yx}}} - \boldsymbol{F}_{\boldsymbol{V}_y} \right)^2 \right). \tag{1}$$

Let $\boldsymbol{C}_c$ represent the centroids that results in minimum reconstruction error, i.e., $\boldsymbol{x} = \boldsymbol{c}$. Then, we can create $\boldsymbol{F}_{\boldsymbol{V}_{yc}}$ by replacing each point in $\boldsymbol{F}_{\boldsymbol{V}_y}$ with its corresponding centroid from $\boldsymbol{C}_c$. Let $\boldsymbol{C}_{cN}$ Be the cluster centroid of the cluster $\mathbf{W_N}$, $\boldsymbol{C}_{\mathrm{cS}}$ for cluster $\mathbf{W_S}$ and $\boldsymbol{L}_{\mathrm{cS}}$ for cluster $\mathbf{W_L}$. These cluster centroids ($\boldsymbol{C}_{cN}$, $\boldsymbol{C}_{\mathrm{cS}}$, $\boldsymbol{L}_{\mathrm{cS}}$) correspond to a subset of hematoxylin, eosin, and background pixels in the training images, respectively. Accordingly, pixel locations corresponding to these cluster centroids in $\boldsymbol{F}_{\boldsymbol{V}_{yc}}$ should correspond to hematoxylin, eosin, and background pixels in the test image

$$\boldsymbol{\Omega} = \{\boldsymbol{\omega} \mid \boldsymbol{F}_{\boldsymbol{V}_{yc}}(\boldsymbol{\omega}) == \boldsymbol{C}_{\mathrm{cN}}\} \tag{2}$$

$$\mathbf{K} = \{\boldsymbol{\kappa} \mid \boldsymbol{F}_{\boldsymbol{V}_{yc}}(\boldsymbol{\kappa}) == \boldsymbol{C}_{\mathrm{cS}}\} \tag{3}$$

$$\boldsymbol{\Gamma} = \{\gamma \mid \boldsymbol{F}_{\boldsymbol{V}_{yc}}(\gamma) == \boldsymbol{C}_{\mathrm{cL}}\} \tag{4}$$

where $\boldsymbol{\Omega}$, K, and $\boldsymbol{\Gamma}$ contain the pixel locations corresponding to subset of hematoxylin, eosin, and background regions in the test image, respectively. Now, one can compute the color deconvolution matrix as

$$\left\langle \boldsymbol{CD} = -\log \frac{\sum^{\boldsymbol{V}_y}(\boldsymbol{\Omega})}{\left|\sum^{\boldsymbol{V}_y}(\boldsymbol{\Omega})\right|} \left| -\log \frac{\sum^{\boldsymbol{V}_y}(\mathrm{K})}{\left|\sum^{\boldsymbol{V}_y}(\mathrm{K})\right|} \right| - \log \frac{\sum^{\boldsymbol{V}_y}(\boldsymbol{\Gamma})}{\left|\sum^{\boldsymbol{V}_y}(\boldsymbol{\Gamma})\right|} \right\rangle^{-1} \tag{5}$$

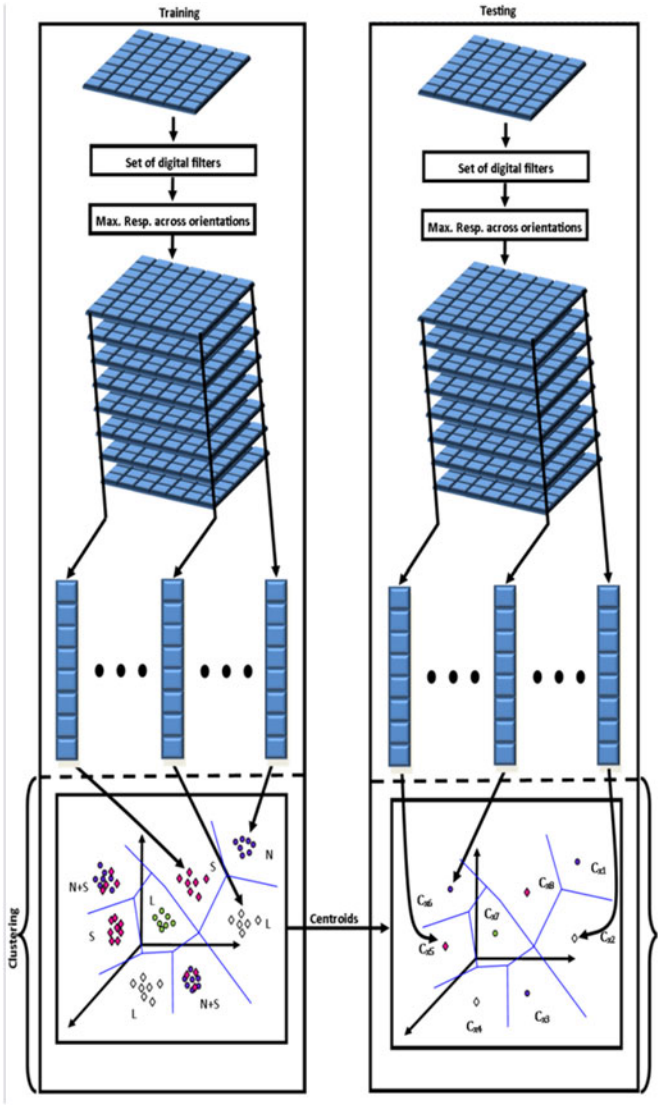


Fig. 1. Block diagram of the segmentation method. The N, S, and L in the block diagram represent nuclei, stroma, and luminal spaces, respectively. During the training, the input image is filtered through a set of filters. The filter set contains 38 different types of filters at multiple orientations. For each filter type, only maximum response across all orientation is recorded. So, for every pixel, this results in its corresponding eight-dimensional feature vector. This eight-dimensional feature vectors are utilized to create clusters. For testing, the whole process is repeated except for clustering. In testing, clustering is replaced by computing the minimum distance between the eight-dimensional feature vectors in the test image to the cluster centroids (resulting from the clusters in the training). Exclusion of clustering from the test phase makes it efficient. Moreover, it provides a straight forward way to establish correspondence between training and test clusters.

and the stain concentration (SC) at each point can be determined by

$$\mathbf{SC} = \boldsymbol{CD} \times (-\log \boldsymbol{T}_{\boldsymbol{x}}). \tag{6}$$

The first channel of $SC$ approximates stain concentration of hematoxylin at each pixel in $T_x$ while the second and third channels approximate the concentration of Eosin at each pixel in $T_x$ and background pixels, respectively. Fig. 2 shows an exam-

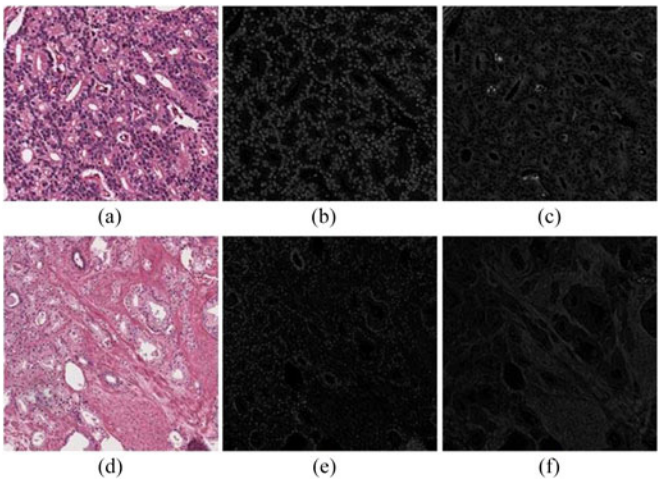


Fig. 2. Examples of color deconvolution with the proposed segmentation method. All images are scanned at 40 × magnification. The image size is 3000 × 3000 pixels. (a) An example of prostate image. (b) Concentration of hematoxylin (a). (c) Concentration of eosin in (c). (d) Another example of prostate image. e) Concentration of hematoxylin in (d). (e) Concentration of eosin in (d). It is obvious that the hematoxylin concentration within nuclei is higher in (b) and (e) as compare to (c) and (f). Moreover, the eosin concentration is higher in (c) and (f) as compared to (b) and (e).

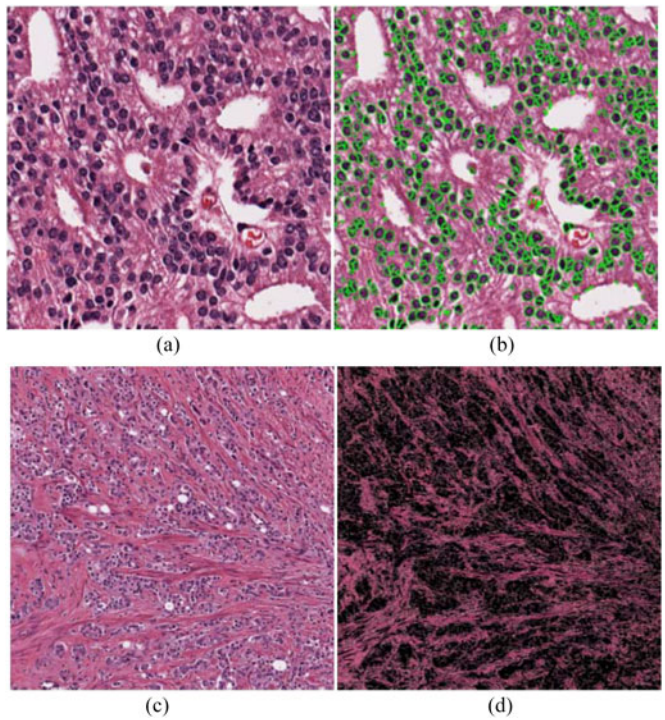


Fig. 3. Example of nuclei detection and stroma segmentation. (a) Original image scanned at 40 × magnification and size 900 × 900 pixels. (b) Nuclei detected by the proposed method. Nuclei are overlaid in green for clarity. (c) ROI example image scanned at 40 × magnification and size $2600 \times 2600$ pixels. (d) Segmented stroma region.

ple of color devolution achieved by the proposed segmentation method.

The SC is further subjected to entropy thresholding [24] to segment nuclei and stroma from $T_x$. To identify the individual nuclei from nuclei clusters, we used normalized multiscale difference of Gaussian as described in [25]; further details of the segmentation method are discussed in [26]. Fig. 3 shows an example of resulting nuclei and stroma.

### B. Lumen, Nuclei, and Stroma Feature Extraction

The first section introduces the methodology to extract novel features related to luminal spaces, while the second section presents a framework to extract architectural changes in nuclei and stroma and use them as features for classification.

*1) Luminal Spaces-Based Features:* The feature extraction method starts by detecting the set of most discriminatory (low vs. high grade) luminal spaces ($L_D$) from the segmented image, $L$. First, we compute the shortest distance of each skeleton to its region border to remove regions whose maximum shortest distance to its border is less then $\delta_h$ (we experimentally set $\delta_h = 20$). Let $L = \{\varsigma_1, \varsigma_2, \cdots, \varsigma_n\}$ represents the set of $n$ potential luminal regions. This can be efficiently computed as

$$\boldsymbol{L_D} = \{\boldsymbol{\varsigma_i} | \mathbf{max}\left(\wp\left(\boldsymbol{\varsigma_i'}\right)\right) > \boldsymbol{\delta_h}\}. \tag{7}$$

Here $\wp$ represents the distance transform [27] and $\varsigma_i{}'$ denotes the complement of the binary region $\varsigma_i$. The maximum distance computed via the distance transform from the complement of a binary region is equivalent to computing the maximum distance from the skeleton of a binary region to its border. Fig. 4(a) shows an example of a prostate image scanned at 40 × magnification. Fig. 4(b) shows the set of potential luminal spaces resulting from the segmentation of the unstained pixels, i.e., the third channel of SC (6). Fig. 4(c) shows the distance transform of the complement of the binary regions from the borders to the skeleton of their respective regions. Fig. 4(d) shows the set of set of discriminatory luminal spaces $L_D$.

Clearly, we do not include small luminal spaces in $L_D$. Due to the huge variation in the shape of luminal spaces, we did not add any further constraints except for the one mentioned in (7). Moreover, it was shown in [28] and [29] that the shape of the luminal spaces has the potential to discriminate between intermediate Gleason grades. So, there is a possibility that adding further constraints might decrease the potential of luminal spaces.

$L_D$ detection is followed by finding the "nearest" epithelial nuclei for each $\varsigma_i$ in $L_D$. Here "nearest" is defined as the innermost layer of epithelial nuclei ($Ep_i$) surrounding $\varsigma_i$. Once $Ep_i$ nuclei for each $\varsigma_i$ are extracted, we compute the percentage of $Ep_i$ to the total number of nuclei in an image

$$\boldsymbol{P_{L_D}} = \left(\sum\nolimits_{\boldsymbol{i}} |\boldsymbol{Ep_i}| / \mathbf{total\ number\ of\ nuclei}\right) \times \mathbf{100}. \tag{8}$$

Our hypothesis is that for low-grade Gleason images, $P_{L_D}$ will be higher. Our hypothesis is mainly driven by two observations:

1) In high-Gleason-score images, the luminal spaces tend to get smaller. Due to their smaller size, (7) will discard them from $L_D$. This will result in a fewer luminal spaces

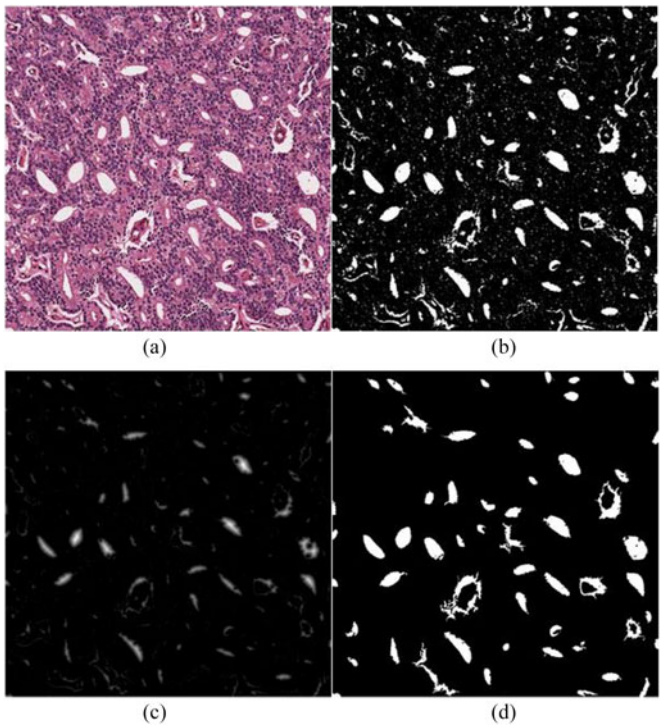


Fig. 4. Finding the set of discriminatory luminal spaces. (a) Prostate image of size 3000 × 3000 pixel scanned at 40 × magnification. (b) Result of stain deconvolution followed by image segmentation. (c) Shortest distance of border pixels to the skeleton computed using (7). (d) $L_D$, set of discriminatory luminal spaces found by using $\delta_h = 20$. It is obvious that $L_D$ only contains large luminal spaces.

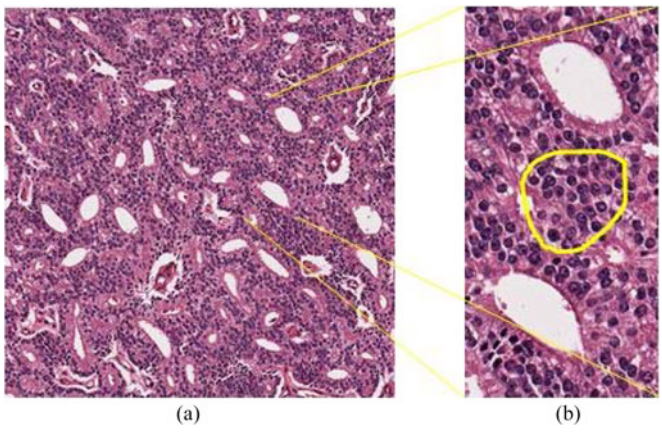


Fig. 5. High-score prostate image. (a) Prostate image of size 3000 × 3000 pixel scanned at 40 × magnification. (b) Zoomed version of Fig. 3.2(a). The yellow outline shows a nuclei cluster formed due to the fusion of lumens. These clusters are common in high-Gleason-score images.

in $L_D$. As a consequence, this will reduce the number of nuclei in $Ep_i$.

2) Stroma is mainly responsible for holding the shape of a tissue. In high-Gleason-score images, the stroma tends to disappear or get considerably small between the lumens, i.e., the lumens tend to fuse together. This gives rise to "nuclei clusters" at the junction of lumen borders. One such cluster is outlined in Fig. 5(b). As a consequence of these clusters, the percentage of $Ep_i$ to the total number of nuclei in the image decreases (7).

Based on these two observations, we believe that $P_{L_D}$ has the potential to differentiate between low and high Gleason scores. Moreover, $P_{L_D}$ is a histopathologically meaningful feature. To extract $Ep_i$, we utilized the framework which we presented in [30]. Briefly, in [30], we presented a gray weighted distance transform to compute the inter-nuclei distances to detect and classify high cell-density regions. Here, we utilized the same framework to compute the shortest distance between $\varsigma_i$ and $Ep_i$. To compute this, let us consider an image $f$ defined over a digital space. An $nD$ (n-dimensional) *digital space* usually refers to an $nD$ grid space that only contains integer points in $nD$ Euclidean space, i.e., $Z^n \subset R^n$, where $Z^n$ represents the n-dimensional discrete space while $R^n$ stands for real coordinate space. An $nD$ digital image $f$, can be defined as a function on $Z^n$. Here, $f$ is a binary image with all pixels set to one except the interior of each nucleus. Here the nuclei image $(SC_2)$ is computed by subjecting the second channel of $SC$ in (6) to entropy thresholding [24]. The interior can be set to zero by using $SC_2 - (SC_2 \ominus SE)$. Here, $SE$ is a 3 × 3 square structuring element and $\ominus$ represents the morphological erosion.

Let $p, q \in Z^n$ and $P_{\mathrm{pq}}$ represent the set of all possible paths between pixels $p$ (border pixels in $\varsigma_i$) and $q$ (border pixel of nuclei). Let $\pi \in P_{\mathrm{pq}}$ represent a path of length $\mathfrak{H}$ as $\mathfrak{H}$-tuple such that

$$\boldsymbol{\pi} = (\boldsymbol{p} = \boldsymbol{p}_0; \boldsymbol{p}_1; \cdots; \boldsymbol{p}_{\mathfrak{H}-1} = \boldsymbol{q}) \tag{9}$$

where $p_i$ and $p_{i+1}$ are adjacent pixels (8 connectivity). Then, the length $\mathrm{Len}$ of the path $\pi$ is defined as

$$\mathrm{Len}(\boldsymbol{\pi}) = \sum_{i=0}^{\mathfrak{H}-1} \boldsymbol{f}(\boldsymbol{p}_i) + \boldsymbol{f}(\boldsymbol{p}_{i+1}) \tag{10}$$

$$= \boldsymbol{f}(\boldsymbol{p}) + \boldsymbol{f}(\boldsymbol{q}) + 2\sum_{i=1}^{\mathfrak{H}-2} \boldsymbol{f}(\boldsymbol{p}_i). \tag{11}$$

Exclusion of spatial weighting in (11) ensures that pixels set to zero do not play a role in distance computation, i.e., the distance should not change within the nucleus boundary. This ensures that the nuclei sizes do not affect the distance computation. Essentially, one can travel within the nuclei without incurring any cost, i.e., two different pixels are separated by null distance if there exists a path with zero values linking the pixels [31]. So, this will favor paths through the nuclei instead of paths outside nuclei. From implementation point of view, $f(q)$ corresponds to the seed image to initialize the distance computation, i.e., the border of the nuclei, while $f(p)$ represents the border of the lumens. Now, the gray-weighted distance between $p$ and $q$ can be defined as a combinatorial optimization problem as [32], [33]

$$\boldsymbol{d}(\boldsymbol{p}, \boldsymbol{q}) = \begin{cases} \min\limits_{\boldsymbol{\pi} \in \boldsymbol{P}_{\mathrm{pq}}} \mathrm{Len}(\boldsymbol{\pi}), & \text{if } \mathrm{Len}(\boldsymbol{\pi}) \neq +\infty \\ +\infty, & \text{otherwise.} \end{cases} \tag{12}$$

This basically corresponds to the shortest distance between pixels $p$ and $q$. Now, for each nuclei, the location with the minimum distance is utilized to initialize the backtracking toward the

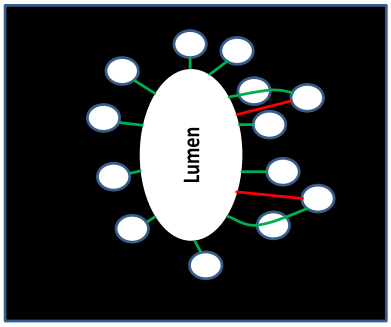


Fig. 6. Schematic diagram of the shortest path of nuclei to the lumen. Here the white circles represent the nuclei. The green line shows the shortest path, while the red line shows few of the longest paths. At first glance, the red paths might look shorter; however, as the cost of traveling through another nucleus is zero, the farthest nuclei will prefer to travel through the other nuclei instead.

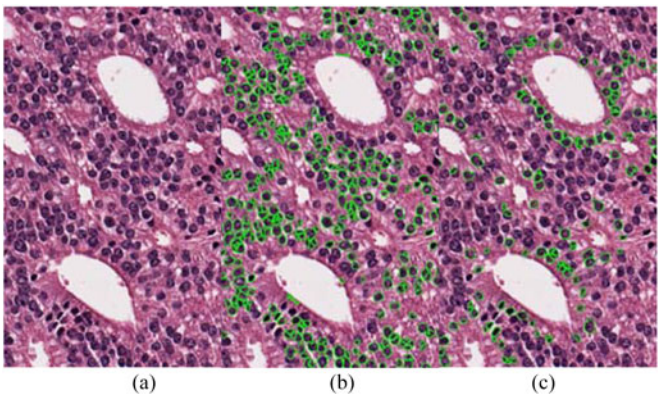


Fig. 7. Example of epithelial nuclei ($Ep_i$) detection. (a) Original image of size 1036 $\times$ 624 pixels. (b) The nonepithelial nuclei overlaid in green. (c) The $Ep_i$ nuclei overlaid in green.

closest point in $\varsigma_i$ [34]. If this path does not contain (crosses) any other nuclei in its path during backtracking, then that nucleus is considered as part of $Ep_i$. Fig. 6 shows the resulting paths as a schematic diagram. Fig. 7(c) shows an example image where all $Ep_i$ are overlaid in green. Once $Ep_i$ is computed, computing $P_{L_D}$ becomes trivial. Moreover, we also use the average shortest distance of $Ep_i$ to its closest nuclei as a proxy to average nuclear to cytoplasmic (NC) ratio. We represent this as $\overline{NC}$.

Using the same framework, we also compute the distance, $D_i^N$ of each nucleus to its closest luminal space. The only difference is in the construction of $f$ which additionally contains $+\infty$ in stroma regions, i.e., cost of traveling through the stroma region is $+\infty$. This prevents the nuclei of one lumen to be assigned to another by creating a natural dam (stroma) between the lumens. Once $D_i^N$ is computed, we take the average of $D_i^N$ and use this as a feature. In low-grade prostate images, this value will be lower due to the presence of well-defined stroma boundaries; however, it will increase in high grade prostate images due to the presence of nuclei clusters at lumen boundaries. The denser the cluster gets, the longer this distance will be.

*2) Measure of Architectural Changes:* In prostate cancer, stroma regions as well as the nuclei have a variable edge (boundary) thickness. We are interested in finding the relationship between "architectural changes of these edges" in different orientations to Gleason grading. Here, orientation needs to be defined in a manner as perceived by the human visual system.

During visual analysis, we do not look for orientation changes at the pixel level; instead, we observe changes at the global level. So, finding the relationship between the architectural changes of edges in different orientations requires the selection of suitable methods for the following: 1) edge detection; 2) edge orientation estimation; and 3) measuring architectural changes.

From a technical point of view, variable edge thickness implies that the edges appear at different scales [35]. Generally, edge detection can be performed either at a fixed or at a variable scale [36]. However, we are interested in filters that can concurrently detect edges occurring at different scales. Sobel, Prewitt, Laplacian [37], etc., are normally used for detecting edges occurring at a fixed scale. However, the filter response is poor at locations where the filter window is smaller or larger than the edge thickness. To overcome this issue, we can apply varying size filters to detect edges occurring at different scales. However, designing a large size filter is a challenging problem. Moreover, due to the local nature of the window of these filters, the orientation estimates are local in nature, and that is not how humans perceive edges in an image.

Gabor filters [38] are good alternatives for detecting edges at different scales. Gabor filters provide directional analysis for particular frequency and orientation, but are local in nature. While Gabor filters provide a mechanism to detect edges occurring at different scales and orientation, how to combine different scales and orientations is an open problem. Decimation free Directional Filter bank (DDFB) appears to be an appropriate candidate as it provides a mechanism to design orientation selective filters irrespective of the scale [36], [38], [39]. These filters split the Fourier transform of an image into wedge shaped passbands. In the spatial domain, these passbands correspond to all features (edges) in a specific orientation irrespective of the scale. Moreover, DDFB detects edges at a global level, i.e., how it is perceived by humans as it is not affected by small changes in orientation.

Let $f_1, f_2, f_3$, and $f_4$ represent the directional filters responsible for extracting edges having orientation $[\frac{\pi}{2}\ \frac{3\pi}{4}]$, $[\frac{\pi}{4}\ \frac{\pi}{2}]$, $[\frac{3\pi}{4}\ \ \pi]$, and$[\pi\ \ \frac{\pi}{4}]$, respectively.

Now to estimate the architectural changes of the edges in different orientations for an image $I$ with size $m \times n$,we compute

$$\Omega_j = \boldsymbol{f}_j \otimes \boldsymbol{I} \tag{13}$$

$$\boldsymbol{C}_j = \frac{\left(\sum_{x=1}^{n}\sum_{y=1}^{=m}|\Omega_j(\boldsymbol{x},\boldsymbol{y})|\right)}{\boldsymbol{m}*\boldsymbol{n}}. \tag{14}$$

Here $j\,\{1,2,3,4\}$ ,and $f_j$ represents the spatial domain directional filters. $C_j$ (stroma) denotes the architectural changes of edges in a certain orientation. To make $C_j$ rotation invariant, we compute

$$\boldsymbol{k} = \max_j(\boldsymbol{C}_j) \text{ with } \boldsymbol{j}\ \epsilon\{1,2,3,4\} \tag{15}$$

$$\boldsymbol{C}'_j = \boldsymbol{C}_{((4-\boldsymbol{k}+\boldsymbol{j})\ \text{modulo}\ 4)+1}. \tag{16}$$

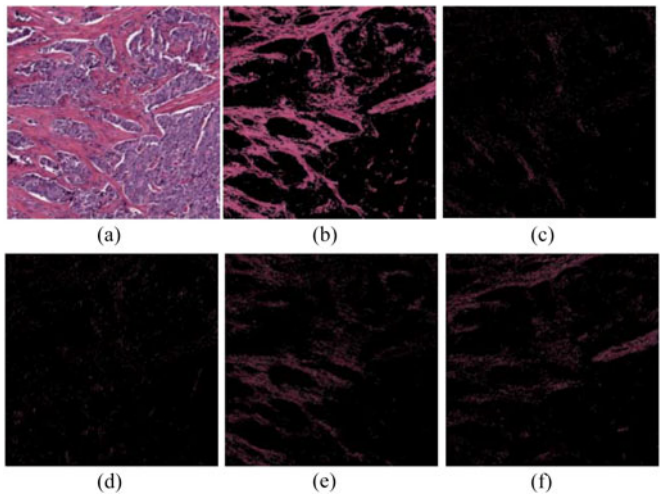


Fig. 8. Example of architectural features of stroma. (a) Original image of size 3000 × 3000 pixels. (b) Extracted stroma region. (c)–(f) The four architectural features proposed in this study. The stroma distribution is not uniform across these images, which according to our hypothesis is associated with high Gleason score.

TABLE I
FEATURES BASED ON THE RELATIONSHIP OF LUMINAL SPACES TO NUCLEI

| Luminal space based features | | |
|---|---|---|
| 1 | $P_{L_D}$ | A scalar representing the ratio of "nearest" epithelial nuclei to all nuclei. |
| 2 | $\overline{D_i^N}$ | A scalar showing the average distance of nuclei to their closest luminal space. |
| 3 | $\overline{NC}$ | Approximation of nuclear to cytoplasmic ratio. |

$C_j'$ in (15) denotes the rotation invariant version of $C_j$. Equations (15) and (16) force the maximum $C_j$ to become $C_1'$ by circularly shifting the vector $\{C_1, C_2, C_3, C_4\}$.For instance, if $C_3$ is maximum among $\{C_1, C_2, C_3, C_4\}$, the application of (15) and (16) will change it to $\{C_3, C_4, C_1, C_2\}$. Fig. 8 shows an example of architectural features extraction from stroma. Fig. 8(c)–(f) shows the four oriented images. As can be noticed, the energy of stroma is non-uniformly distributed across all orientated images. Similarly, we compute $N_j'$, which denotes the architectural changes of nuclei edges in different orientations.

As mentioned earlier, in high-score images the stroma tends to disappear or get considerably small between the lumens. The stroma may also get extraordinarily large in high-score images. So, our hypothesis is that the directional energy will not be uniformly distributed across all directions in high-score images. In case of low-score images, the stroma tends to be uniformly oriented across all directions.

Once the features from Section III-A and Section III-B are extracted, we stack these features, $[P_{L_D}, \overline{D_i^N}, \overline{NC}, C_j', N_j']$, together to create a 11-dimensional feature vector (see Tables I and II for details). Then, we utilize the linear subspace method presented in [44] to classify images into low and high Gleason grade. The linear subspace method is based on singular value decomposition which facilitates to formulate the classification problem as least square minimization problem. It helps us to create two separate linear subspaces (four-dimensional), one for the low grade and the other for the high grade prostate images. The grade associated with a linear subspace, which results in the minimum reconstruction error in terms of Euclidean distance is considered as the prediction for the test image.

TABLE II
SUMMARY OF ARCHITECTURAL FEATURES OF STROMA AND NUCLEI

| Architectural features | | |
|---|---|---|
| 1 | $C_j'$ | A four element vector representing the architectural changes of stroma in four different orientations. |
| 2 | $N_j'$ | A four element vector showing the architectural changes of nuclei in four different directions. |

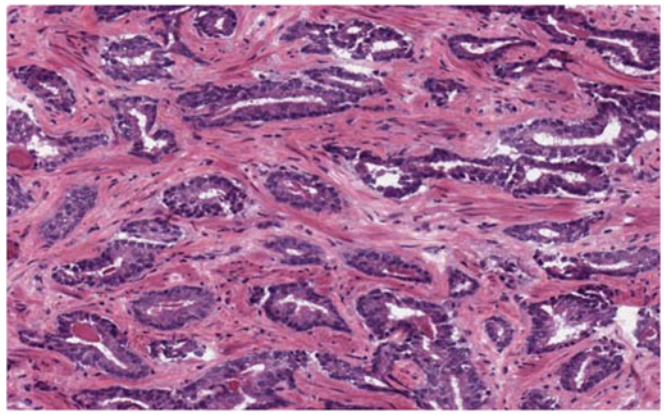

Fig. 9. Low-score ROI image of size 949 × 1325 misclassified during the training. The proposed method failed to detect all the luminal spaces and nuclei in this ROI image.

## III. DATABASES

The importance of histopathological diagnosis of prostate cancer led to the establishment of prostate whole slide (PWS) imaging database within The Cancer Genome Atlas (TCGA) [40]. This TCGA database contains H&E images of prostate patients along with their associated Gleason score. For the training of the proposed method, we used nine PWS images from TCGA database with Gleason sore $\leq 6$ (low score) and 19 PWS images with Gleason score $\geq 8$ (high score). From each of these 28 images (9+19), tumor regions were manually identified and ROIs were extracted by both an expert pathologist and a pathology resident. ROIs were extracted to account for the heterogeneity of the prostate cancer. Most of the PWS images contain two pieces of tissue per slide which allowed us to extract multiple ROIs per image. A total of 43 ROI images were extracted; 14 ROIs from low grade PWS images and 29 from the high grade PWS images. Fig. 9(a) shows one of two tissues from a PWS image. Here, the tumor regions are outlined in green by an expert pathologist. An ROI of size 1889 x 2452 was extracted from the image on the left. Fig. 9(b) shows an ROI image extracted from one of the cancer regions shown in Fig. 9(a).

All ROI images were presented to an expert pathologist for Gleason scoring using an in-house software which provides the functionality to score the images via a web browser.

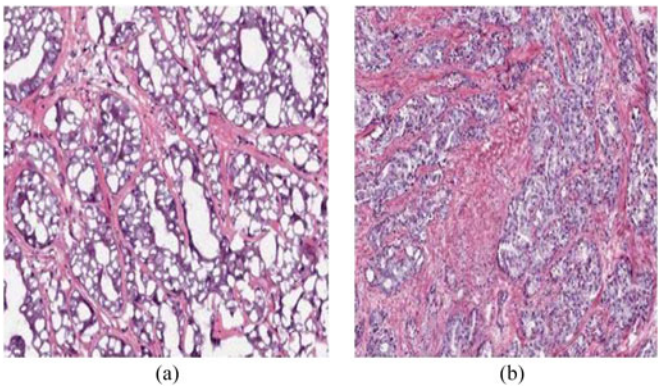


Fig. 10. High-score ROI images misclassified during the training. (a) A high-score ROI image of size $794 \times 1051$. The segmentation method failed to detect the nuclei in this ROI image. (b) An ROI of size $2204 \times 2560$ where the segmentation method failed to detect the nuclei. The aspect ratio of both ROI images was changed for better visualization.

For the validation of the proposed method, we used an independent PWS H&E image database. These images are part of a collection of images obtained from the Physicians' Health Study and the Health Professionals Follow-up Study, a prospective cohort of US male health professionals. These images represent a subset of subjects who developed prostate cancer and underwent prostatectomy [41]. The database consists of 17 low-score and 13 high-score PWS images. Again, tumor regions were manually identified and ROI images were extracted by an expert pathologist and a pathology resident. A total of 88 ROI images were extracted from these 30 PWS images (17+13); 49 from low-score PWS images and 39 from high-score PWS images. Both TCGA and OSUWMC databases were acquired at 20 × magnification. The TCGA and Ohio State University Wexner Medical Center (OSUWMC) ROI images have an average size of 2 K × 2 K and 2.5 K × 2.5 K, respectively.

## IV. Results

For the sake of clarity, we have divided this section into training and test results.

### A. Training Results

The ROIs from TCGA database were used in the training of the proposed method. We used the score assigned by the expert pathologist as the ground truth. Based on leave-one-out-cross validation, the method correctly classified 40 out of the 43 ROI images from the TCGA database. So, the overall training accuracy was 93.0% (40/43). One ROI image (see Fig. 10) from low-score PWS image was misclassified because the proposed method failed to detect all the luminal spaces and nuclei in this ROI image.

The reason for the misclassification of the other two ROIs from the high-score PWS images was a segmentation error; the nuclei could not be segmented properly (see Fig. 10).

Table III shows the results in terms of a confusion matrix. The columns of the confusion matrix represent the predicted score, while the rows represent the Gleason score assigned by the expert pathologist, i.e., the ground truth. The main diagonal of this matrix represent the number of correctly classified images while the anti-diagonal shows the number of miss-classified images.

TABLE III
CONFUSION MATRIX BASED ON THE 43 ROI IMAGES DURING THE TRAINING OF THE PROPOSED METHOD

| ROI Ground Truth \ ROI Predicted Score | Low Score | High Score |
|---|---|---|
| Low Score | 13 | 1 |
| High Score | 2 | 27 |

The columns represent the predicted score, while the rows represent the ground truth. The first row shows that 13 out of 14 low-score ROI images were correctly classified. The second row shows that 27 out of 29 high-score ROI images were correctly classified. The antidiagonal of the confusion matrix shows the number of misclassified ROI images.

TABLE IV
CONFUSION MATRIX BASED ON THE 25 PWS IMAGES DURING THE TRAINING OF THE PROPOSED METHOD

| PWS Ground Truth \ PWS Predicted Score | Low Score | High Score |
|---|---|---|
| Low Score | 6 | 0 |
| High Score | 1 | 18 |

We followed the majority voting rule to assign a score to a PWS image. The main diagonal of this matrix represents the number of correctly classified PWS images, while the antidiagonal shows the number of misclassified PWS images.

One might think that the notion of "visually meaningful features interpretable to the pathologist" is perhaps lost when used in conjunction with a subspace method. To address this issue, we repeated our experiments with random forest, a classifier whose results are relatively simple to interpret. We randomly divided the TCGA database into approximately 10 equal sized partitions. The objective was to choose nine different partitions for training and the remaining partition for validation. During the subsequent folds, it was repeated by considering a different partition for validation and remaining partitions for training. This resulted in an average accuracy of 92.1% with standard deviation of ±0.9258. These results are nearly identical to what we have achieved with the subspace method. From here, one can conclude that with a bit of effort, one can easily interpret which features are responsible for the predicted score. So, for ambiguous cases, the pathologist can actually see the features which are responsible for a certain score. However, for computational efficiency and slightly better performance of the subspace method, we preferred it over random forest.

If we follow the majority voting rule (two or more ROI images from the same PWS image need to be correctly identified) to make a decision, then 24 out of the 25 PWS images were correctly classified, resulting in a training accuracy of 96.0% (24/25). The respective confusion matrix is shown in Table IV.

TABLE V
CONFUSION MATRIX BASED ON THE 25 PWS IMAGES DURING THE TRAINING OF THE PROPOSED METHOD

| PWS Ground Truth | PWS Predicted Score: Low Score | PWS Predicted Score: High Score |
|---|---|---|
| Low Score | 5 | 1 |
| High Score | 2 | 17 |

A PWS image is considered correctly classified if all its respective ROI images are correctly identified.

TABLE VI
CONFUSION MATRIX BASED ON THE 88 ROI IMAGES IN THE OSUWMC DATABASE

| ROI Ground Truth | ROI Predicted Score: Low Score | ROI Predicted Score: High Score |
|---|---|---|
| Low Score | 43 | 6 |
| High Score | 2 | 37 |

However, the training accuracy drops to 88.0% (22/25) if we require all the ROI images from the same PWS image to be correctly identified. Table V shows the results in terms of a confusion matrix.

### B. Test Results

The 88 ROI from the OSUWMC database were used for the evaluation of the proposed method. The method correctly classified 80 out of the 88 ROI images from the OSUWMC database, resulting in an overall testing accuracy of 90.9% (80/88). The resulting confusion matrix is shown in Table VI. As indicated in the confusion matrix, the proposed method misclassified six low and two high score ROI images. It is worth mentioning that none of the images in the OSUWMC database were used during training. This makes our evaluation completely unbiased. Moreover, the OSUWMC database was acquired at a different institution. So, there exist variations in slide preparation and digitization between training and test dataset. The higher classification accuracy on OSUWMC database shows the robustness of the proposed method to variations in slide preparation and digitization.

The ROI images in the OSUWMC database, along with the predicted scores, were presented to the expert pathologist to review the results. Table VII reports the findings of the expert pathologist on the misclassified ROI images.

The expert pathologist considered 6 out of the 8 misclassified ROI images as controversial, and recommended that we discard them from the OSUWMC database. The expert pathologist considered it as a failure of the proposed method in two of the remaining ROI images. This resulted in an overall test accuracy of 97.6%. Table VIII shows the resulting confusion matrix after the thorough analysis of the results.

TABLE VII
THOROUGH EVALUATION OF THE PROPOSED METHOD BY AN EXPERT PATHOLOGIST

| ROI # | GT | PS | Comments after thorough investigation of the results by an expert pathologist | Conclusion |
|---|---|---|---|---|
| 1 | ≤ 6 | ≥ 8 | It is better to call it 3+4 as it contains small, poorly formed glands | Discard |
| 2 | ≤ 6 | ≥ 8 | The tumor is only present in the lower left and lower right corners of the ROI. Still 3+3. The proposed method failed because it included the whole image. | Controversial, discard |
| 3 | ≤ 6 | ≥ 8 | There are no lumens because the lumens are small — this is still 3+3. The method failed due to the small size of the lumen; | Controversial, discard |
| 4 | ≤ 6 | ≥ 8 | it has lumens; there's debris due to the procedure: autolysis (tissue is breaking down). | Controversial, discard |
| 5 | ≤ 6 | ≥ 8 | It contains small glands with small lumens; still 3+3. The lumen is filled with mucin. The method failed to detect luminal spaces. | ≤ 6, Keep |
| 6 | ≤ 6 | ≥ 8 | Small poorly formed glands - some of them have luminal spaces, and some do not. still a Gleason 3+3. Glands are compressed. There's a feature of sclerosis. | Controversial, discard |
| 7 | ≥ 8 | ≤ 6 | It is a 3+4 | Discard |
| 8 | ≥ 8 | ≤ 6 | It is a 4+4; The proposed method failed to segment the nuclei. | ≥ 8, Keep |

Here GT represents the ground truth. PS stands for the predicted score by the proposed method. The conclusion column shows the final recommendations of the pathologist.

TABLE VIII
CONFUSION MATRIX AS A RESULT OF THE REVISED ROI OSUWMC DATABASE

| ROI Ground Truth | ROI Predicted Score: Low Score | ROI Predicted Score: High Score |
|---|---|---|
| Low Score | 43 | 1 |
| High Score | 1 | 37 |

The term "revised" emphasize the fact that the misclassified images were reviewed and analyzed by the expert pathologist as mentioned in Table VII.

TABLE IX
CONFUSION MATRIX RESULTING FROM THE 30 PWS IMAGES IN OSUWMC DATABASE

| PWS Ground Truth | PWS Predicted Score: Low Score | PWS Predicted Score: High Score |
|---|---|---|
| Low Score | 16 | 1 |
| High Score | 1 | 12 |

A PWS image is considered correctly classified if all its respective ROI images are correctly identified.

Based on majority voting, all 30 slides were correctly classified, which resulted in a test accuracy of 100.0% (30/30). However, the training accuracy drops to 93.3% (28/30) if we require all the ROI images from the same PWS image to be correctly identified. The resulting confusion matrix is shown in Table IX.

We have also compared our proposed method with [17]. The method in [17] resulted in an overall accuracy of 79.07% with standard deviation of $\pm 2.45$ on the ROI images in the training database during tenfold cross validation. It misclassified a total of nine images (three from low-score ROI while six from high-score ROI images) while the accuracy dropped to 74.39% for OSUWMC database. A total of 21 images out of 82 were misclassified. We picked [17] for comparison as it is also designed to differentiate between low Gleason score from high. We further compared our results with [42]. For the training database, it resulted in an overall accuracy of 81.65% with standard de-

viation of $\pm 4.79$ during tenfold cross validation. For OSUMC database, the accuracy dropped to 78.32%.

We have used multiple ROI images per patient during leave-one-out cross classification. So, one can think that this might result in over-fitting. However, leave-one-out cross validation was only used during training. The test set (OSUWMC dataset) was completely independent of the training dataset. So, if there had been any over-fitting during training, the method would not have generalized to an independent dataset (the OSUWMC dataset). Leave-one-out for the training set enabled us to make best use of the available data. Moreover, prostate cancer is well known to be heterogeneous, so ROI images from the same patient are relatively different from one another. In fact, leave-one-out cross validation during training shows the *consistency* of the proposed method within same patient.

## V. Discussion

Our contribution in this paper is twofold: 1) identification of visually meaningful architectural and cytological features that can be employed to classify prostate cancer differentiation; and 2) development of a novel method to encode these features for prostate cancer.

The main emphasis of this study is to define a perceptually meaningful objective criterion for prostate cancer grading. The lack of human interpretation is arguably one of the main reasons why current systems have not gained acceptance in the pathology community, although several systems have the required regulatory approval. For instance, in [17], Khurd *et al*. presented a Gleason grading system to differentiate Gleason pattern 3 from 4. They combined textons with support vector machines to classify pixels into Gleason patterns 3 and 4. In [18], Sparks and Madabhushi presented a statistical shape model to perform manifold regularization. They successfully showed that their model could be utilized to differentiate between Gleason patterns 3 and 4. Although these methods reported a high accuracy, the methods are not easily interpretable by the pathologist. The visually meaningful features are designed to provide a pathologist with additional source of information during diagnosis/prognosis as they are motivated by clinical observations.

Comparing the ratio of the number of nuclei ("nearest" epithelial nuclei) within lumens to the total number of nuclei in an image is a feature that can be easily understood by the pathologists. Instead of memorizing some complex and subjective patterns, one can rely on these simple features, which in return provide better discrimination and interpretability. In prostate cancer, the NC ratio is often a useful feature to the pathologist. However, encoding such a feature into an automated image analysis method is challenging as it is difficult to extract a cytoplasmic boundary from H&E images. To overcome this problem, we provide an alternative by computing the shortest distance from the nuclei within a lumen to the border of the luminal spaces. This distance, on average, is smaller in low-score ROI images as compared to high-score ROI images. Indeed, lower Gleason grade patterns such as pattern 3 features discrete and individual gland with minimal distance from the glandular epithelial nuclei to the borders of each gland. Higher grade patterns tend to feature ill-defined or fused glands all the way up to disappearance of glandular architecture, increasing the distance of glandular epithelial cells to the borders of each gland. It is worth mentioning that the proposed method was trained and tested on ROI images. For this reason, it would be illogical to correlate it any other outcome except for Gleason grade. However, the system is general enough to be correlated to any other type of outcome provided that we have access to all slides.

The PWS images often show variations within/between them due to: the type of tissue fixation, embedding, cutting, and staining protocols (such as stain concentration, and staining time) and the image acquisition devices [43], [44]. Some of these variations are very difficult to be addressed via color normalization methods [20], [43], [45] only. These color normalization methods lack explicit explanations on how they tackle these variations. For these reasons, we adopted a variant of textons [22], [46]. Most of our filters for texton generation rely on the edge and curvature information from the red channel of the RGB color image, hence, compensating for most of the color variations. However, the method still failed to segment a few ROI images which points toward the complexity of the variations in the PWS images. Unlike [17], we utilized textons to extract nuclei, stroma, and lumens rather than use them for classifying ROI images into low and high Gleason score.

The training and testing of the proposed method are reported on a reasonable size dataset, i.e., 55 H&E slides. The nearly identical results on two independent databases, TCGA and OSUWMC, indicate the robustness and generalizability of the proposed method. We attribute the success of the method to visually meaningful features generation. We believe that these features can be considered as complimentary features to the existing feature set in the Gleason grading system. Moreover, these features have the potential to be considered as independent features for scoring of WSP images.

Replacing the MR8 filter bank in our segmentation framework with the filter bank presented in [22] produces comparable results; however, we preferred MR8 due to its computational efficiency [46]. It is also worth mentioning that the training takes an average of 5 mins on the set of 5 training images of size $2500 \times 2500$ pixels in MATLAB 2012a. The majority of the time is taken by $k$-means clustering. Fortunately, this can be performed offline. As far as testing is concerned, it takes nearly 10 seconds to perform color deconvolution on a $2500 \times 2500$ pixels RGB color image. In testing, the majority of the time is spent on solving (1). Increasing the number of training images might increase the robustness of the color deconvolution; however, it will increase the computational cost during testing. Due to the parallelizable nature of (1), the computational cost during testing can be overcome by exploiting the parallel architecture of modern day computers.

Although the proposed method performs at a reasonable accuracy, the emphasis of this study was not to replace the pathologists. The reasonable accuracy of the proposed method establishes a new set of visually meaningful features that can differentiate between low and high Gleason score. The objective is to assist the pathologists in comparing and validating their findings with that of the proposed method. A method that allows

the pathologist to have an objective reason for a certain score provides the pathologists with an extra bit of confidence during diagnosis/prognosis, and will enable them to share their findings with peers/residents/students in plain English. Our experience with developing computer-assisted diagnosis systems for other diseases [3], [47]–[50] indicates that pathologists have stronger acceptance and trust for the tools and systems that they can understand. This experience motivated the design of this study.

Dividing grades of Gleason scores into smaller numbers of groups can reduce the variability [5], [6]. The current treatment options for Gleason grades $3+3$ or less (i.e., $1+1, 1+2, 2+1$, etc.) is the same. Similarly, the current treatment options for Gleason grades or $4+4$ or more (i.e., $4+4, 4+5, 5+4$, etc.) is the same. Hence, smaller number of groups for these categories is possible and will reduce variability while not impacting stratification. However, some categories $3+4$ and $4+3$ are quite different and their discrimination is very important. There may even be further subcategories within these two. This further emphasizes the need for objective, repeatable, quantitative methods to grade different prostate cancers. Our effort is in that direction and our future work will extend our feature set to include intermediate Gleason patterns 3 (Gleason score$3+4$) from pattern 4 (Gleason score$4+3$).

## VI. Conclusion

In this paper, we introduced a new set of visually meaningful features to inform a reproducible digital image analysis method to quantify prostate cancer grade. The developed method uses distilled, objective criteria derived from a widely accepted Histopathological method that helps pathologist to grade prostate cancer with better precision and accuracy. In situations where there is a disagreement between the Gleason score assigned by a pathologist and the proposed method, the visually meaningful property allows the pathologist to further scrutinize the results.

In our future work, we will extend this feature set to include intermediate Gleason pattern 3 (Gleason score$3+4$) from pattern 4 (Gleason Score$4+3$). One possibility would be to apply the current method on small patches instead of ROI images and investigating if the method can differentiate between Gleason patterns 3 from 4. Currently, the system requires ROI images from the tumor regions to predict a Gleason score. However, we will investigate the possibility of automatically localizing tumor regions from PWS images. We will further investigate the possibility of including multiple pathologists into our future studies.

## References


[1] D. J. Tindall and P. J. Scardino, "State of research for prostate cancer: Excerpt from the report of the prostate cancer progress review group," *Urology*, vol. 57, pp. 28–30, 2001.

[2] D. F. Gleason and G. T. Mellinger, "The veterans administration cooperative urological research group: Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging," *J. Urol.*, vol. 111, pp. 58–64, 1974.

[3] I. Thompson, J. B. Thrasher, G. Aus, A. L. Burnett, E. D. Canby-Hagino, M. S. Cookson, A. V. D'Amico, R. R. Dmochowski, D. T. Eton, and J. D. Forman, "Guideline for the management of clinically localized prostate cancer: 2007 update," *J. Urol.*, vol. 177, pp. 2106–2131, 2007.

[4] A. J. Evans, P. C. Henry, T. H. Van der Kwast, D. C. Tkachuk, K. Watson, G. A. Lockwood, N. E. Fleshner, C. Cheung, E. C. Belanger, and M. B. Amin, "Interobserver variability between expert urologic pathologists for extraprostatic extension and surgical margin status in radical prostatectomy specimens," *Amer. J. Surgical Pathol.*, vol. 32, pp. 1503–1512, 2008.

[5] G. D. Carlson, C. B. Calvanese, H. Kahane, and J. I. Epstein, "Accuracy of biopsy Gleason scores from a large uropathology laboratory: Use of a diagnostic protocol to minimize observer variability," *Urology*, vol. 51, pp. 525–529, 1998.

[6] 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 Pathol.*, vol. 32, pp. 81–88, 2001.

[7] M. McLean, J. Srigley, D. Banerjee, P. Warde, and Y. Hao, "Interobserver variation in prostate cancer Gleason scoring: Are there implications for the design of clinical trials and treatment strategies?" *Clinical Oncol.*, vol. 9, pp. 222–225, 1997.

[8] T. Y. Chan, A. W. Partin, P. C. Walsh, and J. I. Epstein, "Prognostic significance of Gleason score 3+ 4 versus Gleason score 4+ 3 tumor at radical prostatectomy," *Urology*, vol. 56, pp. 823–827, 2000.

[9] C. Mosquera-Lopez, S. Agaian, A. Velez-Hoyos, and I. Thompson, "Computer-aided prostate cancer diagnosis from digitized histopathology: A review on texture-based systems," *IEEE Rev. Biomed. Eng.*, vol. 8, pp. 1–16, 2014.

[10] 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 Pathol.*, vol. 35, pp. 1121–1131, 2004.

[11] R. Stotzka, R. Männer, P. H. Bartels, and D. Thompson, "A hybrid neural and statistical classifier system for histopathologic grading of prostatic lesions," *Anal. Quant. Cytol. Histol./Int. Acad. Cytology Amer. Soc. Cytol.*, vol. 17, pp. 204–218, 1995.

[12] Y. Smith, G. Zajicek, M. Werman, G. Pizov, and Y. Sherman, "Similarity measurement method for the classification of architecturally differentiated images," *Comput. Biomed. Res.*, vol. 32, pp. 1–12, 1999.

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

[14] 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.

[15] R. Farjam, H. Soltanian-Zadeh, R. A. Zoroofi, and K. Jafari-Khouzani, "Tree-structured grading of pathological images of prostate," in *Proc. SPIE Med. Imag.*, 2005, pp. 840–851.

[16] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," vol. 13{1), pp. 284–287, 2012.

[17] P. Khurd, C. Bahlmann, P. Maday, A. Kamen, S. Gibbs-Strauss, E. M. Genega, and J. V. Frangioni, "Computer-aided Gleason grading of prostate cancer histopathological images using texton forests," in *Proc. IEEE Int. Symp. Biomed. Imag.: From Nano to Macro*, 2010, pp. 636–639.

[18] R. Sparks and A. Madabhushi, "Statistical shape model for manifold regularization: Gleason grading of prostate histology," *Comput. Vision Image Understanding*, vol. 117, pp. 1138–1146, 2013.

[19] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vision*, vol. 62, pp. 61–81, 2005.

[20] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal. Quant. Cytol. Histol.*, vol. 23, pp. 291–299, 2001.

[21] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vision*, vol. 43, pp. 29–44, 2001.

[22] C. Schmid, "Constructing models for content-based image retrieval," *IEEE Comput. Vision Pattern Recog. (CVPR)*, pp. II-39–II-45, 2001.

[23] M. Varma and A. Zisserman, "Classifying images of materials: Achieving viewpoint and illumination independence," in *Proc. Comput. Vision ECCV*, 2002, pp. 255–271.

[24] M. K. K. Niazi, M. Pennell, C. Elkins, J. Hemminger, M. Jin, S. Kirby, H. Kurt, B. Miller, E. Plocharczyk, and R. Roth, "Entropy based quantification of Ki-67 positive cell images and its evaluation by a reader study " in *Proc. SPIE Med. Imag.*, 2013, pp. 867601-1–867601-9.

[25] M. K. K. Niazi, A. A. Satoskar, and M. N. Gurcan, "An automated method for counting cytotoxic T-cells from CD8 stained images of renal biopsies," in *Proc. SPIE Med. Imag.*, 2013, pp. 867606-1–867606-10.
[26] M. K. K. Niazi, J. H. Chung, K. J. Heaton-Johnson, D. Martinez, R. Castellanos, M. S. Irwin, S. R. Master, B. R. Pawel, M. N. Gurcan, and D. A. Weiser, "Advancing clinicopathologic diagnosis of high-risk neuroblastoma using computerized image analysis and proteomic profiling," *Pediatric and Developmental Pathology*, pp. 1–8, 2017.
[27] G. Borgefors, "Distance transformations in digital images," *Comput. Vis. Graph. Image Process.*, vol. 34, pp. 344–371, 1986.
[28] S. Naik, S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "Gland segmentation and computerized Gleason grading of prostate histology by integrating low-, high-level and domain specific information," in *MIAAB (Microscopic Image Analysis with Applications in Biology) workshop*, pp. 1–8, 2007.
[29] S. Naik, A. Madabhushi, J. Tomaszeweski, and M. D. Feldman, "A quantitative exploration of efficacy of gland morphology in prostate cancer grading," in *IEEE 33rd Annu. Northeast Bio. Eng. Conf.*, pp. 58–59, 2007.
[30] M. K. K. Niazi, G. Beamer, and M. N. Gurcan, "Detecting and characterizing cellular responses to Mycobacterium tuberculosis from histology slides," *J. Cytometry Part A*, vol. 85, pp. 162–168, 2014.
[31] P. Soille, "Generalized geodesic distances applied to interpolation and shape description," in *Mathematical Morphology and Its Applications to Image Processing*. New York, NY, USA: Springer, 1994, pp. 193–200.
[32] M. Gedda, "Contributions to 3D image analysis using discrete methods and fuzzy techniques: With focus on images from cryo-electron tomography," Ph.D. dissertation, Centre Image Anal., Uppsala Univ., Acta Universitatis Upsaliensis, 2010.
[33] M. Niazi, "Image filtering methods for biomedical applications," Ph.D. dissertation, Centre Image Anal., Uppsala Univ., Acta Universitatis Upsaliensis, 2011.
[34] L. Ikonen and P. Toivanen, "Shortest routes between sets on gray-level surfaces," *Pattern Recognit. Image Anal.*, vol. 15, pp. 195–198, 2005.
[35] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629–639, Jul. 1990.
[36] M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. Image Process.,* vol. 14, no. 12, pp. 2091–2106, Dec. 2005.
[37] S. E. Umbaugh, *Computer Vision and Image Processing: A Practical Approach Using Cviptools with Cdrom*. Englewood Cliffs, NJ, USA: Prentice-Hall PTR, 1997.
[38] M. A. U. Khan, M. K. Khan, and M. A. Khan, "Coronary angiogram image enhancement using decimation-free directional filter banks," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. V-441–V-444, 2004.
[39] M. K. Khan, M. F. Nilsson, B. R. Danielsson, and E. Bengtsson, "Fully automatic heart beat rate determination in digital video recordings of rat embryos," *Advances in Mass Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry*, Springer Berlin Heidelberg, pp. 27–37, 2008.
[40] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, G. M. Mastrogianakis, J. J. Olson, T. Mikkelsen, N. Lehman, and K. Aldape, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–1068, 2008.
[41] S. Lindström, F. R. Schumacher, D. Cox, R. C. Travis, D. Albanes, N. E. Allen, G. Andriole, S. I. Berndt, H. Boeing, and H. B. Bueno-de-Mesquita, "Common genetic variants in prostate cancer risk prediction—Results from the NCI Breast and Prostate Cancer Cohort Consortium (BPC3)," *Cancer Epidemiol. Biomarkers Prevention*, vol. 21, pp. 437–444, 2012.
[42] PW Huang and CH Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE Trans. Med. Imag.*, vol. 28, no. 7, pp 1037–1050, Jul. 2009.
[43] D. Onder, S. Zengin, and S. Sarioglu, "A review on color normalization and color deconvolution methods in histopathology," *Appl. Immunohistochem. Molec. Morphol.*, vol. 22(10), pp. 713–719, 2014.
[44] P. A. Bautista, N. Hashimoto, and Y. Yagi, "Color standardization in whole slide imaging using a color calibration slide," *J. Pathol. Informat.*, vol. 5:4, 2014.
[45] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2009, pp. 1107–1110.
[46] J. M. Geusebroek, A. W. M. Smeulders, and J. Van De Weijer, "Fast anisotropic Gauss filtering," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 938–943, Aug. 2003.
[47] K. Belkacem-Boussaid, S. Samsi, G. Lozanski, and M. N. Gurcan, "Automatic detection of follicular regions in H&E images using iterative shape index," *Comput. Med. Imag. Graph*, vol. 35, pp. 592–602, 2011.
[48] K. Belkacem-Boussaid, M. Pennell, G. Lozanski, A. Shana'ah, and M. Gurcan, "Computer-aided classification of centroblast cells in follicular lymphoma," *Anal. Quant. Cytol. Histol.*, vol. 32, pp. 254–260, Oct. 2010.
[49] S. S. Samsi, A. K. Krishnamurthy, M. Groseclose, R. M. Caprioli, G. Lozanski, and M. N. Gurcan, "Imaging mass spectrometry analysis for follicular lymphoma grading," in *Proc. IEEE Eng. Med. Biol. Soc.*, 2009, vol. 2009, pp. 6969–6972.
[50] M. Gurcan, O. Sertel, J. Kong, A. Ruiz, M. Ujaldon, U. PhD, G. Lozanski, H. Shimada, and J. Saltz, "Computer-assisted histopathology: Experience with neuroblastoma and follicular lymphoma," presented at the Workshop Bioimage Informat. Biol. Imag., Comput. Vision Data Mining 2008, Santa Barbara, CA, USA..

Authors' photographs and biographies not available at the time of publication.