Journal of
Biomedical Optics

BiomedicalOptics.SPIEDigitalLibrary.org

Automatic Gleason grading of prostate cancer using quantitative phase imaging and machine learning

Tan H. Nguyen
Shamira Sridharan
Virgilia Macias
Andre Kajdacsy-Balla
Jonathan Melamed
Minh N. Do
Gabriel Popescu

SPIE.

Tan H. Nguyen, Shamira Sridharan, Virgilia Macias, Andre Kajdacsy-Balla, Jonathan Melamed, Minh N. Do, Gabriel Popescu, "Automatic Gleason grading of prostate cancer using quantitative phase imaging and machine learning," *J. Biomed. Opt.* **22**(3), 036015 (2017), doi: 10.1117/1.JBO.22.3.036015.

Automatic Gleason grading of prostate cancer using quantitative phase imaging and machine learning

Tan H. Nguyen,[a,†] Shamira Sridharan,[a,†] Virgilia Macias,[b] Andre Kajdacsy-Balla,[b] Jonathan Melamed,[c] Minh N. Do,[d] and Gabriel Popescu[a,*]
[a]University of Illinois, Beckman Institute for Advanced Science and Technology, Department of Electrical and Computer Engineering, Quantitative Light Imaging Laboratory, Urbana–Champaign, Illinois, United States
[b]University of Illinois, Department of Pathology, Chicago, Illinois, United States
[c]New York University, School of Medicine, Department of Pathology, New York, New York, United States
[d]University of Illinois, Department of Electrical and Computer Engineering, Computational Imaging Group, Coordinated Science Laboratory, Urbana–Champaign, Illinois, United States

Abstract. We present an approach for automatic diagnosis of tissue biopsies. Our methodology consists of a quantitative phase imaging tissue scanner and machine learning algorithms to process these data. We illustrate the performance by automatic Gleason grading of prostate specimens. The imaging system operates on the principle of interferometry and, as a result, reports on the nanoscale architecture of the unlabeled specimen. We use these data to train a random forest classifier to learn textural behaviors of prostate samples and classify each pixel in the image into different classes. Automatic diagnosis results were computed from the segmented regions. By combining morphological features with quantitative information from the glands and stroma, logistic regression was used to discriminate regions with Gleason grade 3 versus grade 4 cancer in prostatectomy tissue. The overall accuracy of this classification derived from a receiver operating curve was 82%, which is in the range of human error when interobserver variability is considered. We anticipate that our approach will provide a clinically objective and quantitative metric for Gleason grading, allowing us to corroborate results across instruments and laboratories and feed the computer algorithms for improved accuracy. [DOI: 10.1117/1.JBO.22.3.036015]

Keywords: quantitative phase imaging; prostate cancer diagnosis; machine learning; microscopy; holography.

Paper 160878PR received Jan. 12, 2017; accepted for publication Mar. 13, 2017; published online Mar. 30, 2017.

1 Introduction

Prostate cancer is the second leading cause of cancer-related death among men in the United States,[1,2] after lung cancer. In 2015, 220,800 men were diagnosed with prostate cancer, accounting for 26% of the total number of new cancer cases, and 27,540 men are projected to eventually die from the disease.[2] Prostate health is evaluated using different formats, including a detailed medical interview, a physical examination with digital rectal examination (DRE), or a prostate-specific antigen (PSA) blood test. Abnormal DRE results or PSA levels above the normal value of 4 ng/ml might lead to a prostate biopsy to confirm whether these abnormalities are due to cancer.[3] The excised tissue samples are fixed using formalin and then embedded in paraffin wax, which is sectioned into thin slices using a microtome. These sections are then deparaffinized and stained with hematoxylin and eosin (H&E) dye for microscopic examination by the pathologist. If the pathologist suspects the presence of cancer, based on the absence of the myoepithelial or basal cell layer, cancer severity is assessed using the Gleason grading system.[4,5] The Gleason score is the sum of two Gleason grades corresponding to the two most prominent disease patterns present in the examined tissue. The Gleason grade, which typically ranges from 3 to 5, measures the degree of glandular separation and, thus, cancer aggressiveness. The glands in Gleason grade 3 carcinoma are smaller and more closely packed than in a normal prostate, resulting in a reduced separation between them. In Gleason grade 4, the glands display fusion, sometimes creating what appears to be large glands containing multiple lumens, also known as the “cribriform” pattern. In Gleason grade 5, glands are very poorly differentiated with sheets of epithelial cells seen in the stroma, which is connected with poor disease outcome. Although the Gleason grading system has undergone a few revisions since it was first established, it continues to remain a strong prognostic indicator. The Gleason score is linked to several clinical endpoints, including progression to metastatic disease and patient survival.[6] It also influences the treatment decisions made by the physician.[7] Accurate discrimination between Gleason grade 3 and 4 is critical as it triggers the switch between active surveillance and aggressive treatment.[8]

Although the diagnosis of prostate biopsies by a trained pathologist is currently considered to be the “gold standard,” the technique suffers from several shortcomings. First, for Gleason grading, the samples are stained using H&E, aiming to target different components in the prostate biopsies, e.g., nuclei, cytoplasm, and nucleoli. The protein-rich regions, basic in nature, are stained pink while those that are acid rich become blue. Other markers with better specificity have also been developed.[9–11] The need for using these markers stems from the fact that many biopsies are nearly transparent under bright-field microscopy inspection. Therefore, exogenous factors must

*Address all correspondence to: Gabriel Popescu, E-mail: gpopescu@illinois.edu

†These authors have equal contribution.

1083-3668/2017/$25.00 © 2017 SPIE

be introduced to enhance the contrast. This process takes time, requires expertise, and sacrifices the intrinsic properties of the sample. Furthermore, the staining poses a significant challenge for improving the throughput of the system using modern computing algorithms. An experienced pathologist can handle the variation in the concentration of the dye, staining skill, and color balance. However, it requires additional processing and assumptions before inputting to a computer to automate the process. Significant effort has been spent in producing reliable automatic Gleason grading based on the histological H&E images. Such efforts can be divided into two categories: classification- and segmentation-based techniques. Methods in the first group use various features from stained images to produce Gleason scores without the need for image segmentation. These features include textures from H&E images[12] and multispectra images.[13] Methods in the second group produce a Gleason score in two stages. In the first stage, label maps of the biopsies are produced from the H&E images. Then, morphological features are extracted from these maps. Finally, subsequent classifiers are deployed to produce the final Gleason grade. Naik et al.[14] built statistical models of the likelihood for the class of a pixel given its color and location in the training set. Nguyen et al.[15] used the (L, a, b) color space and various constraints on the relative arrangements of tissue regions sections to refine the segmentation map. To achieve automatic histology using H&E images, preanalytical variables, such as exposure time, magnification, illumination spectra, dye concentration, must be normalized to produce color consistency.[16,17] Also, there is a lack of a universal agreement on how the normalization should be performed and what the correct normalization result should be.[18]

Understanding these obstacles, several groups have tried to do diagnosis from label-free slices. Muller et al.[19] used the optical attenuation coefficient measured using needle-based optical coherence tomography as a tool for detection of prostate cancer. They showed that the optical attenuation coefficient was significantly higher in malignant tissue compared to benign prostate tissue. Uttam et al.[20] used optical path length information to quantify the depth-resolved density alteration of the nuclear architecture as a tool for early prediction of cancer progression. Spectroscopy methods have also been used to examine the biochemical information of the tissue at a molecular level for different pathologies *in vitro*. Crow et al.[21] demonstrated the use of Raman spectroscopy to differentiate between benign samples, benign prostatic hyperplasia (BPH) and prostatitis, from prostate cancer at an accuracy of 86%. Combining Fourier transform infrared (FTIR) spectroscopy with bright-field microscopy, Kwak et al.[22] improved the accuracy of automatic segmentation and demonstrated an area under the curve (AUC) of at least 0.97 in a binary classification problem between cancer versus noncancer cases. It was later shown that FTIR spectroscopy can be used to provide a better prediction of prostate cancer recurrence, compared to two widely used tools, Kattan nomogram and CAPRA-S.[23] However, the spectroscopic information in FTIR is obtained at the expense of spatial resolution (typically above 10 to 15 μm) and extremely slow acquisition speed. Fehr et al.[24] suggested using magnetic resonant imaging as a noninvasive tool for automatic classification of Gleason scores.

Recently, quantitative phase imaging (QPI)[25–31] has emerged as valuable tool for rendering high contrast of unlabeled transparent samples. The contrast of QPI is due to the real part of the refractive index of the sample retrieved through interferometric settings. Therefore, the measurement is very robust to change in the illumination condition, e.g., illuminating variation, allowing high repeatability and seamless translation across measurement sites. Previously, many QPI methods utilized laser illumination due to a requirement for long coherence length in "traditional" interferometry. The laser illumination generates random speckle pattern,[32] which suppresses structural details of the biopsy. Recently, a combination between white-light illumination and "common-path" interferometry[26–28,33] has solved this problem. This method, referred to as spatial light interference microscopy (SLIM),[27] allows the refractive index information to be captured at a diffraction-limited resolution with nanoscale accuracy and excellent temporal stability. In Ref. 34, it was reported, for the first time, the potential of QPI to classify cancerous areas versus benign ones in prostate biopsies, using the mean and median of the phase distribution. Furthermore, light scattering parameters measured in the prostate stroma using QPI have been used to predict the aggressiveness of intermediate grade prostate cancer.[35]

Here, we introduce a combination of advanced machine learning algorithms with SLIM label-free imaging and describe the first label-free tissue scanner with automatic prostate cancer diagnosis. The SLIM system can image at 12.5 SLIM images per second with 40× magnification and 4 megapixels/frame. Using a tissue microarray (TMA), containing more than 300 cores, we segmented different regions from prostatectomy samples into multiple classes (gland, stroma, and lumen) with high accuracy. Segmented label maps are further used to obtain several morphological features of the glands of the cores, e.g., distortion, variation of gland areas, etc. One of our advantages over other techniques is the ability to extract physics-related features, e.g., stroma anisotropy, which characterizes the directional dependence of the light scattering when it propagates through stromal areas of the tissue. Using these features, we were able to separate regions with Gleason grade 3 and Gleason grade 4 with an AUC of 0.82.

2 Material and Methods

2.1 *Label-Free Tissue Scanner*

Let $T = e^{i\phi}$ be the transmission of the biopsy and ϕ be the optical phase shift introduced by the sample. We have $\phi = h\Delta n(2\pi/\lambda)$. Here, λ is the central wavelength of 552 μm, Δn is the refractive index difference, and h is the sample thickness. Our system measures ϕ using a commercial inverted phase contrast microscope (Axio Observer, Z1, Zeiss Inc.) connected to an external SLIM module (CellVista SLIM Pro, Phi Optics, Inc.). A schematic of our optical setup is shown in Fig. 1(a). Under uniformly coherent illumination, the imaging field at the output port of the microscope is a magnified version of the transmission, i.e., $U_t = T$. This total field is Fourier transformed into $\tilde{U}_t$ by the lens L_1 onto the surface of a spatial light modulator (SLM) (Boulder Nonlinear Inc.). On this surface, the spatial spectrum $\tilde{U}_t$ is spatially separated into two different components, the nonscattering component, $\tilde{U}_o$, and the scattering one, $\tilde{U}_s = \tilde{U}_t - \tilde{U}_o$. The nonscattering component $\tilde{U}_o$ matches the support of the condenser phase annulus and the phase ring of the phase contrast objective. Meanwhile, the scattering component $\tilde{U}_s$ covers the rest of the aperture of the objective. These Fourier components are inversely Fourier transformed by a secondary lens L_2 into two fields, the nonscattering field, U_o, and the scattering field, U_s. The field U_o spatially averages the total field U_t into $U_o = \langle U_t \rangle_{\mathbf{r}}$. The scattering field, U_s, on the other

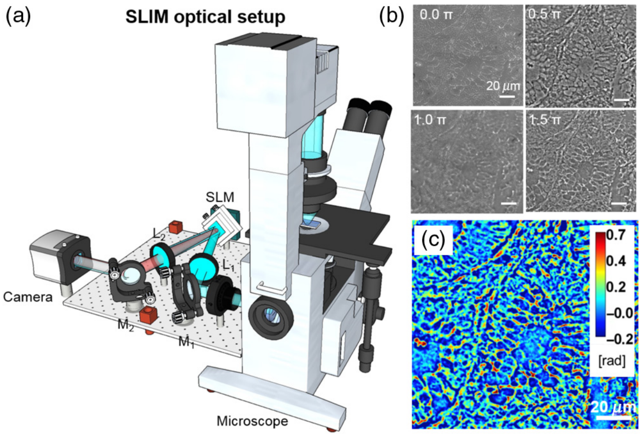

Fig. 1 Optical setup and working principle: (a) SLIM optical setup, (b) four intensity images and the phase map computed by combining four frames, and (c) the resulting SLIM image.

hand, is given as $U_s = U_t - U_o$. The SLM further retards U_o by $n\pi/2$, with $n = 0,1,2,3$, resulting in four interference intensities on the camera plane of

$$\begin{aligned}I(n\pi/2) &= |e^{in\pi/2}U_o + U_s|^2\\ &= |U_o|^2 + |U_s|^2 + 2|U_oU_s|\cos(\Delta\phi - n\pi/2). \quad (1)\end{aligned}$$

These four intensities are captured by a high-resolution 10-bit 5.5-megapixel sCMOS camera (Andor Inc.). The phase is extracted using the following procedures. First, we compute the phase difference between two fields as $\Delta\phi(\mathbf{r}) = \arg(U_s) - \arg(U_o) = \arg\{[I(0) - I(\pi)] + i[I(\pi/2) - I(-\pi/2)]\}$. Here, i is the imaginary unit. The ratio between the amplitude of these two fields $\beta(\mathbf{r}) = |U_s|(\mathbf{r})/|U_o|(\mathbf{r})$ is obtained from the following relations: $[I(0) + I(\pi)]/2 = |U_o|^2 + |U_s|^2$ and $[I(0) - I(\pi)]/4 = |U_o||U_s|$. See Ref. 36 for more details. Finally, the phase of the total field is calculated from β and $\Delta\phi$ as

$$\phi = \arctan\{[\beta \sin \Delta\phi]/[1 + \beta \cos \Delta\phi]\}, \quad (2)$$

where we dropped the coordinate $\mathbf{r}$ for brevity.

Figure 1(b) shows four intensity images for a small section of a prostate biopsy. The first inset is the original phase contrast image at zero external phase modulation while the last inset is the bright-field image with a phase modulation of $3\pi/2$. Figure 1(c) is the extracted phase map for ϕ from these intensities. All details in the intensity images are still visible in the SLIM image with very good contrast. More importantly, the SLIM image is not susceptible to variation in the illumination, as it only captures intrinsic information of the sample.

2.2 *Imaging the Tissue Microarray*

The TMA is provided by the Co-operative Prostate Cancer Tissue Resource from the College of Medicine of University of Illinois at Chicago. It consists of 368 prostate cores (one core per patient) with several diagnosis results, including normal, BPH, high-grade prostatic intraepithelial neoplasia (HGPIN), and Gleason scores varying from $2 + 2$ to $5 + 5$. After being deparaffinized, unstained cores were first imaged using SLIM under a 40× magnification. The phase images are stitched together to generate one high-resolution image per core. Each such image has $10{,}000 \times 10{,}000$ pixels with a pixel ratio of 14 pixels per micron. Afterward, the cores were stained with H&E and scanned by a bright-field tissue scanner. The results of all cores can be found in Fig. 2(a). Figures 2(b) and 2(c) are a zoomed-in H&E image of a core and its corresponding SLIM image, respectively.

Figure 3 displays three H&E [Fig. 3(a)] and SLIM [Fig. 3(b)] images of three cores in the TMA. To provide ground truth for automatic diagnosis, different regions of interest (ROIs) in each core are studied and color-coded by certified pathologists for the diagnosis results using H&E images. In Fig. 3, green color indicates normal areas, red color indicates HGPIN, and blue color indicates tumors. Based on these annotations, corresponding regions in the SLIM images are extracted and further used for the automatic diagnosis.

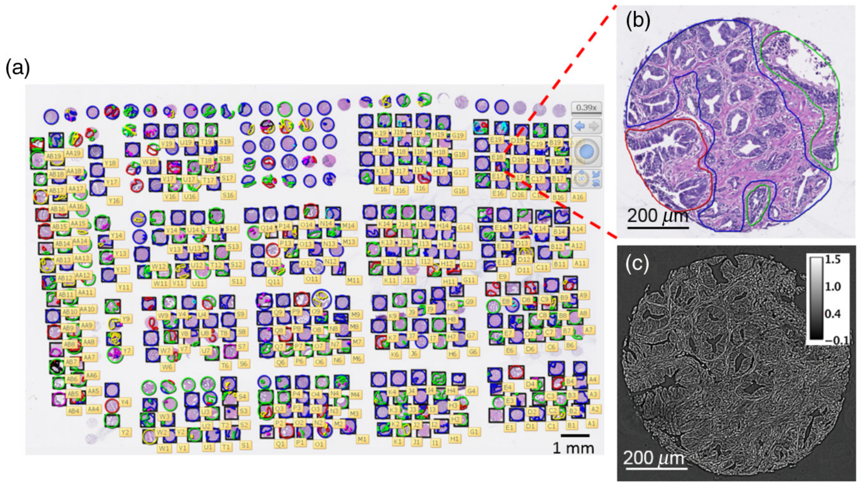

Fig. 2 H&E image of the whole TMA with diagnosis results. (a) H&E image of the whole TMA slide consisting of 368 cores. (b) A zoomed-in H&E image of a prostate core with annotations. The region highlighted in green represents normal glands, region in blue is Gleason grade 3 prostate cancer glands, and region in red corresponds HGPIN. (c) A zoomed-in SLIM image of the same core as in (b), obtained prior to staining. Morphological features in the H&E image are recapitulated by SLIM.

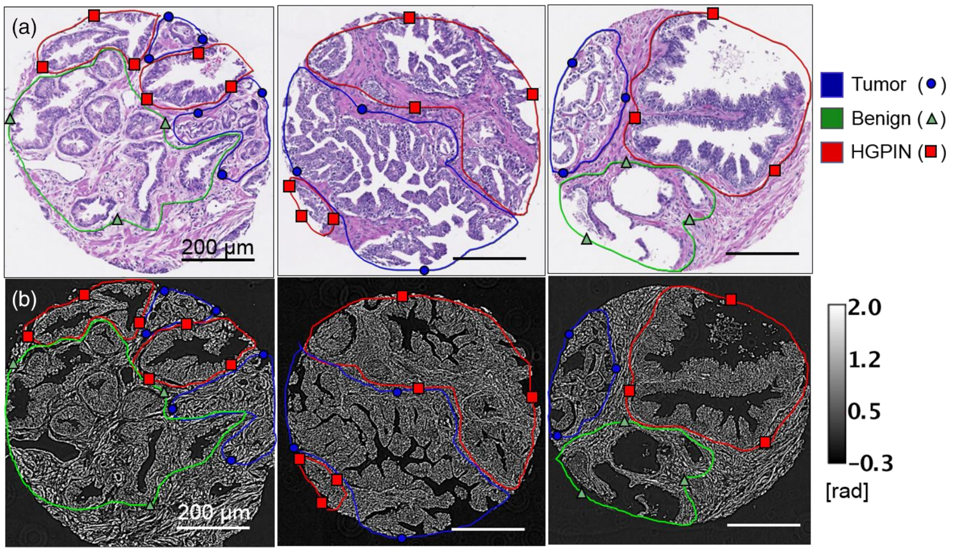

Fig. 3 H&E images versus SLIM images. (a) H&E images of three cores in the TMA. Each core in the training data set includes an annotation of diagnosis results. (b) Corresponding SLIM images of those in (a).

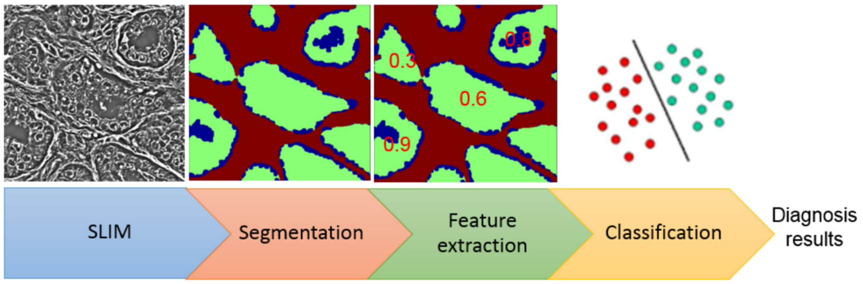

Fig. 4 Automatic diagnosis scheme of steps from an SLIM image to a diagnosis result.

2.3 *Automatic Diagnosis Framework*

To obtain the automatic diagnosis from SLIM images, we use an approach summarized in Fig. 4. First, texture-based features are extracted for each pixel in the SLIM image and passed into a classifier to do automatic segmentation based on pixel classification. Each pixel is assigned into one of three following classes (gland, stroma, and lumen). Second, using the label map obtained from the previous step, morphological and phase-based features are evaluated for all glands in the current field of view and its surrounding stroma. These features are later passed into a subsequent classifier to produce diagnosis results, as described below.

2.4 *Feature Extraction from Phase Images*

To capture the texture expression of the biopsies, we use the "texton" framework proposed by Julesz[37] and later expanded by Leung and Malik.[38,39] The framework has demonstrated great success in solving several computer vision problems, e.g., material classification and characterization,[38–41] due to its ability to accurately imitate mechanism of human's textural perception.[37,42,43] In this work, the framework is used to train a texton dictionary [see Fig. 5(a)] from a set of training images and extract a feature vector for each pixel as follows.

First, each SLIM image ϕ is convolved with a Leung–Malik filter bank,[43] consisting of L filters $h_1, \ldots, h_L$ to generate L filter responses $\phi * h_1, \ldots, \phi * h_L$. The bank consists of 90 filters, $L = 90$, with 10 symmetric ones and 80 directional ones, oriented at eight different angles over five scales. The directional filters are generated from the first- and second-order derivative of Gaussian kernels with an elongation factor of 3. After this step, each pixel has a feature vector of dimension 90. A phase image in the training set generates totally 3072×3072 feature vectors. To reduce the size of the training set, a subset is formed from four million feature vectors randomly selected from a larger pool of all feature vectors, accounting for 0.18% of the total number of feature vectors. Finally, the K-mean clustering algorithm with $K = 50$ is applied on this subset to divide it into K clusters whose centers are chosen as textons. Here, the value of K is chosen to balance between the complexity of the model and the estimation error, i.e., avoiding cases where there is not enough textons to capture texture variation or those when some textons come from clusters only associated to a very small number of filter responses.

In feature extraction, given an input image (I), one wants to obtain a set of descriptors for each pixel in the image. Figure 5(b) shows how to calculate these descriptors. First, filter responses of the image I to the Leung–Malik filter bank are computed. Then, we apply vector quantization to associate the filter

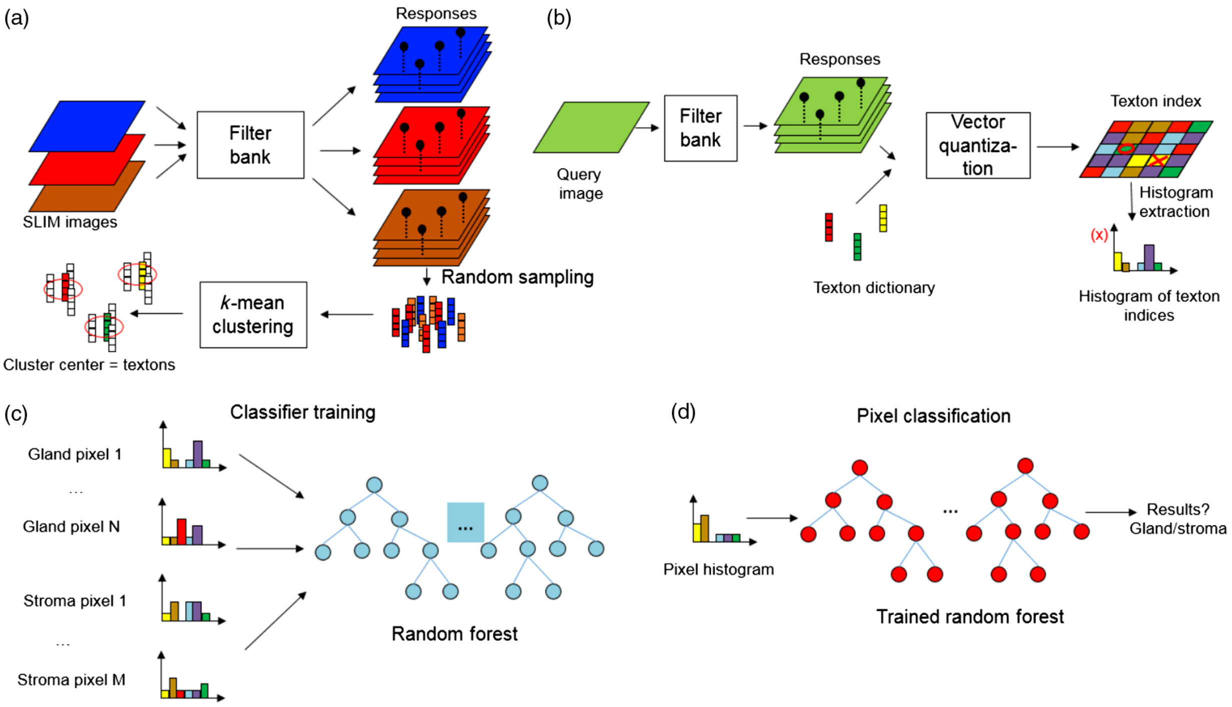

Fig. 5 Feature extraction from SLIM images: (a) extracting the texton dictionary from a training set of SLIM images, (b) feature extraction using vector quantization and the trained dictionary of textons from (a), (c) RF training using descriptors obtained from (b), and (d) pixel classification using a trained RF from (c).

response at each pixel to the closest texton in the dictionary obtained in the previous step. For each pixel i, let us use t_i to denote the index of the closest texton. By definition, t_i can take one of the K following indices $\{1,2,\dots,K\}$. The output of this step is an indexing map where each pixel is assigned to a number, the index of the closest texton. Using this indexing map, for each pixel i, we further obtain a histogram of texton indices evaluated over pixels in its neighborhood. To control the trade-off between the richness of texture information and locality of the descriptor, we apply a Gaussian weight to the histogram calculation where larger histogram contribution is given to pixels closer to the center of the neighborhood. By trial and error, we determine that a neighborhood radius of $\sigma =$ 45 pixels is suitable to characterize the pixel. This radius corresponds to $\sim$10 μm in the downsampled SLIM image.

2.5 *Random Forest for Automatic Segmentation*

To perform image segmentation, we use a random forest (RF) classifier, a method introduced by Breiman[44,45] and Shotton et al.[46] This classifier has shown success in several problems, such as object segmentation,[40,47,48] human-pose estimation,[46,49] and medical image analysis using magnetic resonant imaging,[50–52] due to its ability to reduce the tree dependence with "feature bagging" and "bootstrap" sampling, i.e., random sampling with replacement. In this work, we use the RF to classify each pixel in the image into one of three classes, i.e., epithelial gland, connecting stroma, and lumen. Lumen pixels are classified first based on the proximity of their phase values to that of the background. Then, remaining pixels are classified into either gland or stroma. Here, we train an "extremely randomized" forest[53] of 50 trees. Our implementation is written in MATLAB® with the MexOpenCV wrapper that allows us to call OpenCV routines. The wrapper is obtained from Ref. 54. The training set consists of 4.92 million histograms with two possible labels ("gland" or "stroma"). These histograms of texton are randomly sampled from a larger pool of 2.2 billion histograms. Each tree in the forests is trained on 11 randomly selected features (out of 50, the total number of textons). For each feature of interest, 100 possible thresholds are considered between the minimum and maximum of the feature values for splitting. Let us use S_n to denote the training data set of samples reaching node n'th. At this note, this set is partitioned into a left set, S_l, and a right one, S_r, based on an optimal selected feature ν_n and its optimal threshold t_n, namely $S_l = \{s \in S_n | \nu_n < t_n\}$ and $S_r = S_n \setminus S_l$. Here, (ν_n, t_n) are chosen to maximize the expected gain of information on category, i.e.,

$$(\nu_n, t_n) = \arg\max_{\nu,t}[IG(\nu,t)] \\ = H(S_n) - \frac{|S_l|}{|S_n|}H(S_l) - \frac{|S_r|}{|S_n|}H(S_r), \tag{3}$$

where $IG(\nu,t)$ is the information gain at the current node when the threshold t and the feature ν are used. $H(S_n)$ is the entropy of the set S_n, measuring its degree of class inhomogeneity. The training is recursive and terminated at a leaf node when the maximum depth of 25 levels is reached or less than 20 training histograms left. The training takes ~5 to 6 h. After the training is completed, each leaf node l_m of the m'th tree contains two class likelihood values, $p_{l_m}(\text{gland})$, $p_{l_m}(\text{stroma})$ telling how many training histograms reached it are of gland and stroma pixels, respectively. These probability quantities are used as the confidence value of the classifier produced by the m'th tree when its leaf node is reached.

After the RF has been trained, automatic segmentation results can be obtained by classifying each pixel in the query image to either gland or stroma using its pixel descriptor [Fig. 5(d)]. The optimal class g^* for each pixel is determined by summing the class likelihood values over all the trees and picking the class that maximizes the combined score, i.e.,

$$g^* = \arg \max_{g} \left\{ \sum_{m=1}^{T} p_{l_m}(g) \right\}, \tag{4}$$

where $g \in \{\text{gland}, \text{stroma}\}$. After an initial classification for all pixels of the input image, a postprocessing step is applied to fine tune the segmentation result, making it more stable to gland fusion, which is more popular in high-grade cancer. This step is detailed in Sec. 2.6.

2.6 *Postprocessing of Segmentation Results*

The output of the pixel classifier is usually noisy. We use the following procedure to obtain good segmentation results and resolve the glandular fusion, which is more frequent in high-grade prostate carcinoma.

1. Setting aside all lumen pixels from the segmentation. Assign all remaining pixels with stroma likelihood less than 0.5 to gland pixels. An example of the stroma likelihood map is shown in Fig. 6(a).
2. Denoising gland map. First, we remove all small areas inside and between the glands that have less than 2000 pixels. Next, we perform an opening morphological operation using a “disk” structure element with a radius of 20 pixels. Finally, we remove all glands that have less than 5000 pixels. The result of this step is shown in Fig. 6(b). It also can be seen that there are still several glandular regions detected as big blobs due to small stroma likelihood of the connecting stroma.
3. Watershed segmenting on denoised gland map. This step cuts small joining regions between the glands. First, the denoised gland map is inverted to obtain a nongland map [Fig. 6(c)]. Then, a distance transform is applied to the nongland map to calculate the distance between each pixel in the nongland map to the nearest nonzero pixel [Fig. 6(d)]. Next, the watershed transform[55] is applied on the inverted distance map to obtain an oversegmentation result of the gland map into multiple regions [Fig. 6(e)]. Separating lines between neighboring regions is computed by subtracting the gland map [Fig. 6(b)] from the watershed segmentation result. Then, a closing transform is applied to the map of separating lines to make sure their widths are at least 15 pixels. Figure 6(f) shows the result of this step.
4. Evaluating the separation and recombine glands if needed. This step evaluates each separating line produced by step 3. Here, the mean value of the stroma likelihood is evaluated over each separation line. Lines with mean value of stroma likelihood more than 0.2 are retained. Otherwise, they are eliminated and glands separated by them are rejoined. The step gives a refined gland map [Fig. 6(g)]. Compared to Fig 6(b), the map has resolved several separated glands that were incorrectly fused by a simple thresholding. From the refined map, we obtain the final segmentation result in Fig. 6(h).

3 Results

3.1 *Automatic Segmentation*

Figure 7 shows automatic segmentation results overlaid with the SLIM images. It can be seen that the label map has a good correlation with the H&E images. Figure 8 shows other segmentation examples with H&E and SLIM images of increasing Gleason grade from 3 to 5. Their automatic segmentation results from SLIM images are further shown in Fig. 9.

To quantify the performance of our segmentation, we summarize in Fig. 10(a) receiver operating characteristic (ROC) curves for the segmentation of different diagnosis groups using

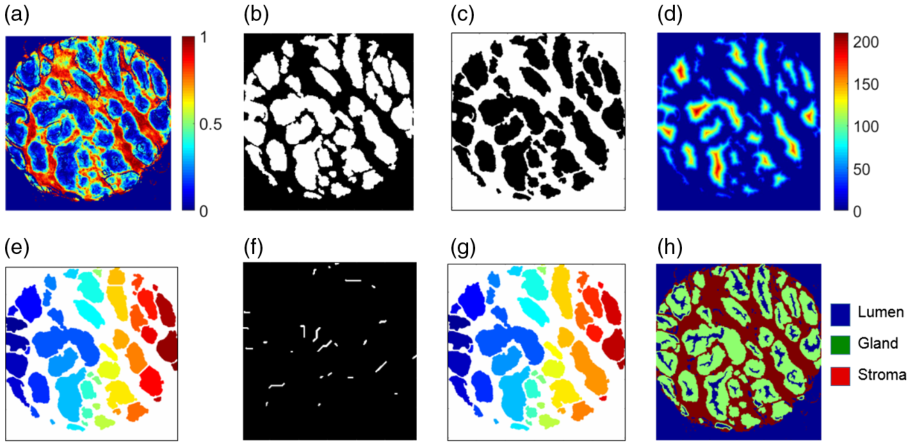

Fig. 6 Steps of image postprocessing for automatic segmentation: (a) the stroma likelihood map produced by the RF classifier, (b) the raw binary gland map, (c) the raw binary nongland map (in white), (d) the distance map from each gland pixel to the nearest nongland pixel, (e) the oversegmentation map produced by watershed segmentation, (f) the map of gland separating lines, (g) the refined segmentation map where some similar regions in (e) have been merged, and (h) the final segmentation map.

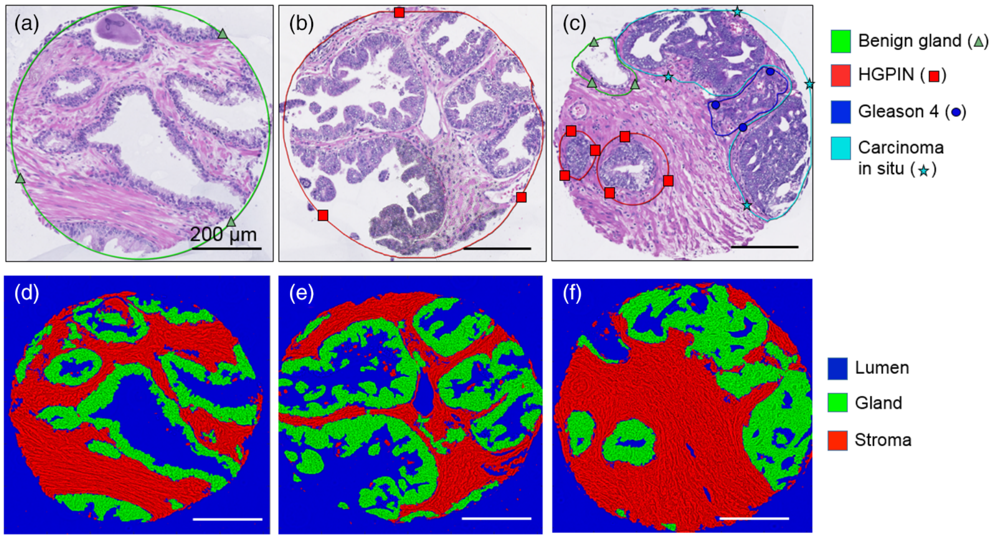

Fig. 7 H&E versus automatic segmentation: (a)–(c) H&E images of three cores and (d)–(f) corresponding automatic segmentation results overlaid on the top of the SLIM images.

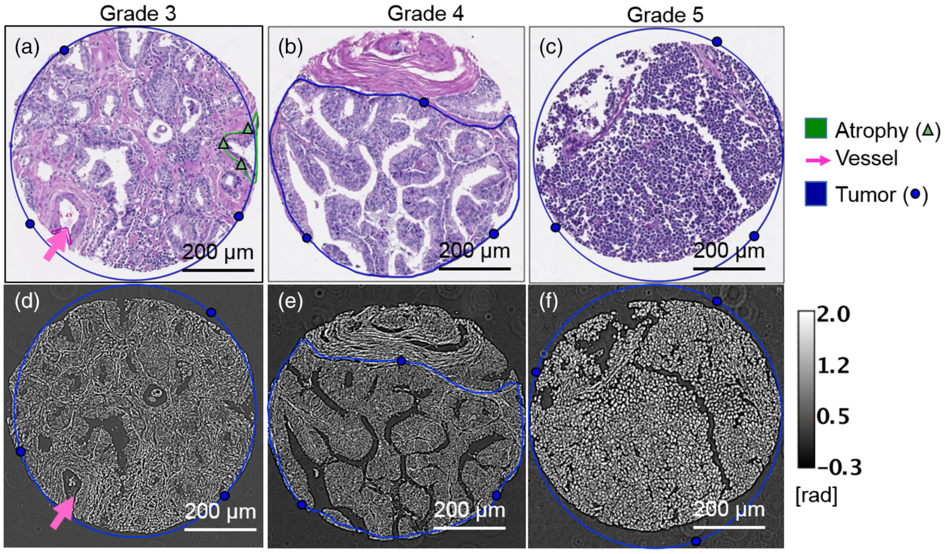

Fig. 8 H&E versus SLIM: (a)–(c) H&E images of three cores in the TMA that have Gleason grade of 3, 4, and 5. (d)–(f) Corresponding SLIM images of these cores. Grayscale bar represents phase shift in radians.

10-fold cross-validation. Since these ROC curves are very close to each other, it is difficult to compare their performance. To address this issue, we show in Fig. 10(b) a zoomed-in image to the upper-left corner of Fig. 10(a), where the true positive rate and the false negative rate range from [0.75, 1.0] and [0, 0.5], respectively. Figure 10(c) shows the corresponding AUC values for each group. The ground truth for the ROC evaluation is created by manually labeling glandular and stromal regions directly on the SLIM images after validating them with H&E images. In the noncancer cases, the AUC is at least 0.97, which indicates that gland and stroma pixels are classified with high accuracy. Meanwhile, in the malignant cases, as cancer progresses to higher grades, e.g., from $3 + 3$ to $5 + 5$, the AUC reduces from 0.98 to 0.87. This result can be explained by the fact that more glandular distortions and deformations are observed at higher grades, which leads to a reduction in discrimination between stroma and glands. Therefore, it is not surprising that the classifier has the smallest AUC with a Gleason score of $5 + 5$. However, these high grades are very easily diagnosed by the pathologist and, thus, do not represent our main focus. Using these segmentation results, we solve the automatic Gleason grading problem, with particular emphasis on discriminating between grades 3 and 4.

3.2 *Automatic Gleason Grading*

Next, we demonstrate the use of our technique in clinical applications with automatic Gleason grading. To generate the ground truth for the training, all cores were reviewed, manually marked, and graded unanimously by two trained pathologists. Figure 2 shows an example of markup results for all cores: 129 regions with Gleason grade 3, 92 regions with grade 4, and 75 regions

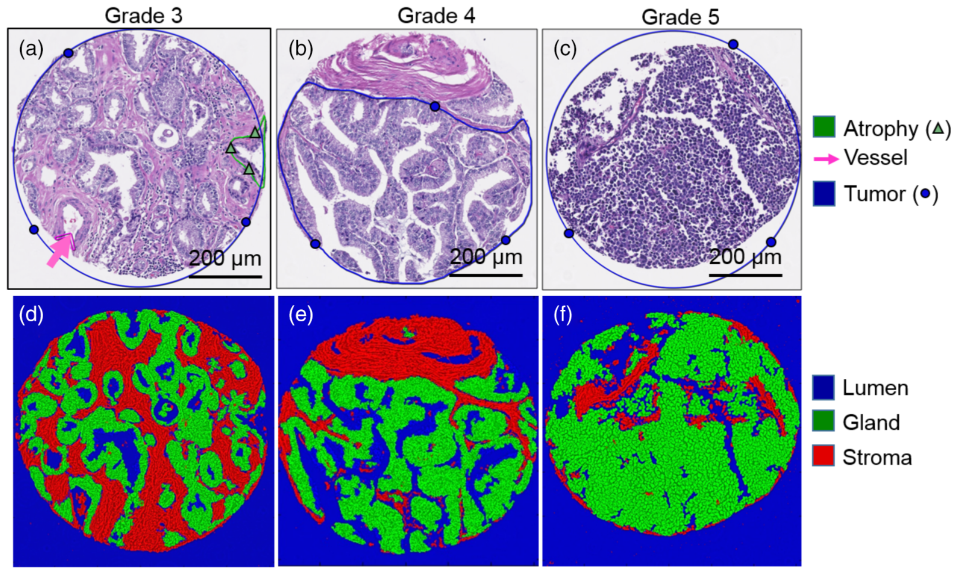

Fig. 9 H&E versus automatic segmentation: (a)–(c) H&E images of three cores of different Gleason grades in Fig. 7, as indicated. (d)–(f) Automatic segmentation results of the cores overlaid on the SLIM images.

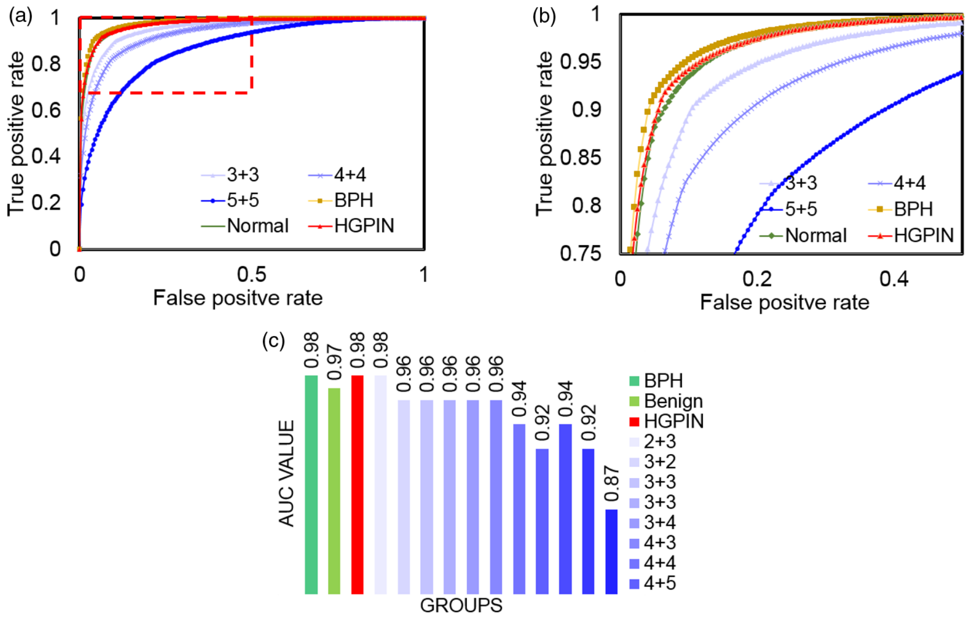

Fig. 10 Automatic segmentation performance. (a) ROC curves for classifying gland versus nongland pixels for all diagnosis groups. The accuracy reduces in the case of higher Gleason grade due to the less defined glands. (b) Zoomed-in plot for a red-dashed region of (a). (c) Bar plot of AUC values of the ROC curves for all diagnosis groups.

with grade 5. Since Gleason grades 2 and 5 are rarely diagnosed, we study the automatic diagnosis problem of differentiating Gleason grade 3 versus 4. It has been shown by Allsbrook et al.[56] that grading 3 versus 4 has a reproducibility problem due to interobserver variation. Also, there is the crucial turning point from active surveillance to aggressive treatment when moving from Gleason grade 3 to 4. Distinguishing between 2 and 5 is not a problem of clinical relevance since it is quite trivial. For example, the authors report an experiment where 38 biopsies with known "consensus" Gleason grade were sent to 41 pathologists to measure interobserver variability. The result was that Gleason grade 4 was undergraded by 21%. Furthermore, there was consistent undergrading of Gleason scores of 5 to 6 (47%), 7 (47%), and 8 to 10 (25%). Clearly, a computer-driven, unbiased procedure for grading is a potential way to tackle this challenge.

Figure 11 shows five different types of features extracted from each region. These features include the mean of glandular distortion, the fusing ratio of glands, the mean number of lumen areas per gland, the coefficient of variation for gland variation,

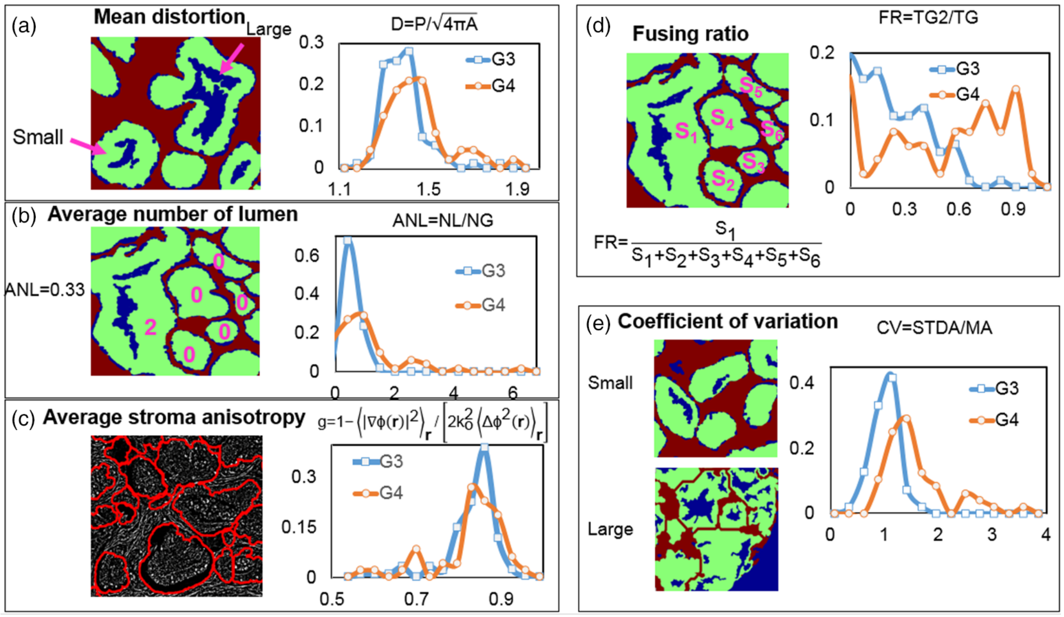

Fig. 11 Feature extraction for automatic diagnosis. Each subfigure shows a feature, how it is calculated and the distributions of the feature values for G3 (blue) and G4 (orange). (a) Mean distortion feature: *D*, mean distortion of a gland; *P*, perimeter of a gland; and *A*, area of a gland. (b) Average number of lumens: ANL, average number of lumens; NL, number of lumen; and NG, number of glands. (c) Average stroma anisotropy. (d) Fusing ratio: FR, fusing ratio; TG2, total areas of glands with at least two lumens; and TG, total area of all glands in current field of view. (e) Coefficient of variation: CV, coefficient of variation; STDA, standard deviation of areas of all glands; and MA, mean of the areas of all glands.

and the mean of stroma anisotropy. Other features include the maximum number of lumen areas, the median of the glandular distortion, median of stroma anisotropy, and the mean circularity. Some features are explained here in detail.

- Gland distortion [Fig. 11(a)]: The distortion of a gland is defined as the ratio between its perimeter and the square root of its area, scaled by the factor of $1/\sqrt{2\pi}$. The smaller values of the distortion correspond to more circular glands. Elongated glands have distortion values larger than 1. A distortion value of 1 is obtained for a circular gland. Average distortion value for the whole ROI is evaluated by averaging distortion values of glands inside the ROI.
- The average number of lumen [Fig. 11(b)]: This feature is designed to capture the cribriform pattern that characterizes grade 4. In grade 4, glands fuse to each other, create larger glands with multiple lumen areas contributed by those of individual glands. A value of zero is given if a gland has no area. A fractional number for the average count of lumens over an ROI is also possible.
- Average stroma anisotropy (g) [Fig. 11(c)]: This feature captures potential effects of reactive stroma in the progression of tumorigenesis.[57,58] This quantity measures the angular uniformity of forward scattering in light–tissue interaction. A lower value of anisotropy means more isotropic scattering and vice versa. The anisotropic factor is computed for all pixels in the image using the scattering-phase theorem as[59] $g = 1 - (\langle|\nabla\phi(\mathbf{r})|^2\rangle_{\mathbf{r}}/2k_o^2\langle\Delta\phi^2(\mathbf{r})\rangle_{\mathbf{r}})$. After the anisotropic factor is obtained for all pixels in the image, the stroma anisotropy for the whole field of view is computed as an average of the quantity, evaluating over a thin layer surrounding segmented glands, as shown in Fig. 11(c). The thickness of this thin layer is chosen to be around 10 μm.
- The fusing ratio [Fig. 11(d)]: This feature is the ratio of total area of all glands with at least two lumens to the total area of all glands in the ROI. It characterizes the cribriform pattern and gland fusion. However, it is more robust than the average number of lumen feature since fused glands with small areas have less impact than those with large areas.
- The coefficient of variation of gland area [Fig. 11(e)]: This is the ratio of the standard deviation of all gland area to the mean of all gland area. The ratio is smaller for ROIs that have more uniformity in gland areas, a criterion for Gleason grade 3.

These features are designed to measure the distortion, area homogeneity of the glands, and the amount of gland fusion. They are often used by pathologists for Gleason grading.[5,6,60] Each ROI is characterized by a feature vector of 11 elements. Diagnosis grades from pathologists are used as ground truth for automatic diagnosis. After computing the feature vectors for ROIs, we use logistic regression to classify the grade 3 ROIs versus grade 4. This classifier models the conditional probability of the Gleason grade (G) given the feature vector $\mathbf{X}$ as $p(G = G_3|\mathbf{X}) = \exp(\boldsymbol{\beta}^T\mathbf{X})/[1 + \exp(\boldsymbol{\beta}^T\mathbf{X})]$, and $p(G = G_4|\mathbf{X}) = 1/[1 + \exp(\boldsymbol{\beta}^T\mathbf{X})]$. Here, $\boldsymbol{\beta}$ is a coefficient vector estimated from the data by maximizing the log-likelihood function of the observations of the training set; see Ref. 61 for more details.

The performance of the logistic regression classifier is shown in Fig. 12 in terms of the ROC using leave-one-out

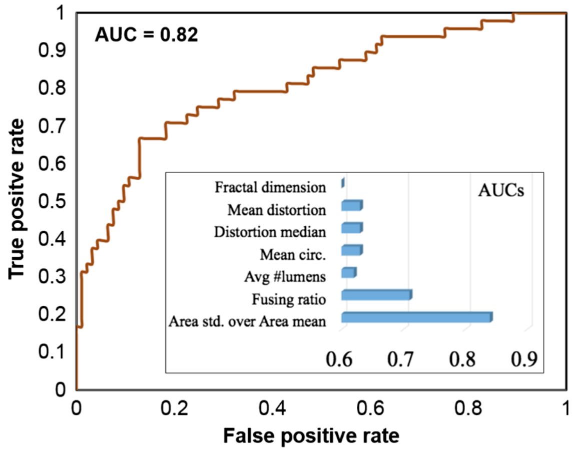

Fig. 12 ROC curves for classifying regions with Gleason G3 versus G4 under leave-one-out cross-validation. These features are extracted from all glands within the ROIs. Diagnosis results from pathologists are used as ground truths. The inset shows AUCs of classification using individual features.

cross-validation. This type of validation trains classifiers using all samples in the dataset except one. A remaining sample is used for testing. The classifier provides two conditional probabilities $p(G_t = G_4|\mathbf{X}_t)$ and $p(G_t = G_3|\mathbf{X}_t)$ for the testing sample X_t. Here, G_t is the Gleason grade of the test sample. Then, we alternate the selection of the testing sample to make sure every sample in our dataset becomes a testing sample once. Finally, the conditional probabilities $p(G_t = G_3|\mathbf{X}_t)$ and $p(G_t = G_4|\mathbf{X}_t)$ of all samples in the dataset are combined to produce the ROC curve for Gleason grade 3 and 4 classification. This curve has an AUC value of 0.82. Note that this error is well within that for interobserver variability reported by Allsbrook et al.[44] The reason why this ROC curve has a stair-case response while the ROC curves shown in Figs. 10(a) and 10(b) are smooth is because of the difference in the number of samples used. To obtain the ROC curve in Fig. 12, we used a dataset consisting of 141 samples. This number is much smaller than that of texture descriptors, about 3.9 millions, used for segmentation in Fig. 10. Since the number of samples is large, the steps of these ROC curve are very small. This fact makes them look smooth. It can be also seen from the curve that, to detect Gleason grade 4 at an accuracy of 90%, the false positive rate will be ∼60%. The inset presents a horizontal bar plot of AUC values when the classifier is trained separately on each individual feature, also using leave-one-out cross-validation. The two largest AUC values are obtained on the coefficient of variation for the areas of the glands and the fusing ratio. Our results demonstrate that label-free imaging and machine learning can provide an objective alternative to pathology, even in the case of difficult tasks, such as classifying Gleason grade 3 and 4.

4 Conclusion

In summary, we have introduced an approach to automatically obtain Gleason grades using QPI and machine learning. Our method combines the merits of QPI, which is insensitive to variation in illumination condition, staining procedure, color balance, etc. Therefore, it allows easy translation across different clinics. The grading process is done automatically, using state-of-the art computational tools to produce objective results and avoid interobserver variation. Here, we use automatic classification of Gleason grade 3 and 4 as an example. The work uses a dataset of 288 cores from a TMA that consists of 368 cores, with consensus diagnosis results. In the future, we aim to validate our algorithm on a larger dataset to make our algorithm more robust to sample variation and to improve the diagnosis accuracy.

Disclosures

G.P. has financial interest in Phi Optics, Inc., a company developing QPI technology for materials and life science applications.

Acknowledgments

This work was supported by the National Science Foundation (No. IIP-1353368) and Agilent Laboratories. For more information, please visit http://light.ece.illinois.edu/.

References

1. U. C. S. W. Group, *United States Cancer Statistics: 1999–2010 Incidence and Mortality Web-Based Report*, Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute, Atlanta (2013).
2. R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA Cancer J. Clin.* **65**(1), 5–29 (2015).
3. S. Frankel et al., "Screening for prostate cancer," *Lancet* **361**(9363), 1122–1128 (2003).
4. D. F. Gleason and E. M. Tannenbaum, "The veteran's administration cooperative urologic research group: histologic grading and clinical staging of prostatic carcinoma," in *Urologic Pathology: The Prostate*, M. Tannenbaum, Ed., pp. 171–198, Lea and Febiger, Philadelphia (1977).
5. J. I. Epstein, "An update of the Gleason grading system," *J. Urol.* **183**(2), 433–440 (2010).
6. P. A. Humphrey, "Gleason grading and prognostic factors in carcinoma of the prostate," *Mod. Pathol.* **17**(3), 292–306 (2004).
7. M. S. Cookson et al., "Correlation between Gleason score of needle biopsy and radical prostatectomy specimen: accuracy and clinical implications," *J. Urol.* **157**(2), 559–562 (1997).
8. H. B. Carter et al., "Gleason score 6 adenocarcinoma: should it be labeled as cancer?" *J. Clin. Oncol.* **30**(35), 4294–4296 (2012).
9. S. Signoretti et al., "p63 is a prostate basal cell marker and is required for prostate development," *Am. J. Pathol.* **157**(6), 1769–1775 (2000).
10. Z. Jiang et al., "P504S/α-methylacyl-CoA racemase: a useful marker for diagnosis of small foci of prostatic carcinoma on needle biopsy," *Am. J. Surg. Pathol.* **26**(9), 1169–1174 (2002).
11. I. L. Deras et al., "PCA3: a molecular urine assay for predicting prostate biopsy outcome," *J. Urol.* **179**(4), 1587–1592 (2008).
12. J. Diamond et al., "The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia," *Hum. Pathol.* **35**(9), 1121–1131 (2004).
13. R. Farjam et al., "Tree-structured grading of pathological images of prostate," *Proc. SPIE* **5747**, 840–851 (2005).
14. S. Naik et al., "Gland segmentation and computerized Gleason grading of prostate histology by integrating low-, high-level and domain specific information," in *Proc. of 2nd Workshop on Microsopic Image Analysis with Applications in Biology*, Piscataway (2007).
15. K. Nguyen, B. Sabata, and A. K. Jain, "Prostate cancer grading: gland segmentation and structural features," *Pattern Recognit. Lett.* **33**(7), 951–961 (2012).
16. P. A. Bautista, N. Hashimoto, and Y. Yagi, "Color standardization in whole slide imaging using a color calibration slide," *J. Pathol. Inf.* **5**(1), 4 (2014).
17. Y. Yagi, "Color standardization and optimization in whole slide imaging," *Diagn. Pathol.* **6**(1), S15 (2011).
18. M. N. Gurcan et al., "Histopathological image analysis: a review," *IEEE Rev. Biomed. Eng.* **2**, 147–171 (2009).

19. B. G. Muller et al., "Prostate cancer diagnosis by optical coherence tomography: first results from a needle based optical platform for tissue sampling," *J. Biophotonics* **9**(5), 490–498 (2016).
20. S. Uttam et al., "Early prediction of cancer progression by depth-resolved nanoscale mapping of nuclear architecture from unstained tissue specimens," *Cancer Res.* **75**(22), 4718–4727 (2015).
21. P. Crow et al., "The use of Raman spectroscopy to identify and grade prostatic adenocarcinoma in vitro," *Br. J. Cancer* **89**(1), 106–108 (2003).
22. J. T. Kwak et al., "Multimodal microscopy for automated histologic analysis of prostate cancer," *BMC Cancer* **11**(1), 62 (2011).
23. J. T. Kwak et al., "Improving prediction of prostate cancer recurrence using chemical imaging," *Sci. Rep.* **5**, 8758 (2015).
24. D. Fehr et al., "Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images," *Proc. Natl. Acad. Sci. U. S. A.* **112**(46), E6265 (2015).
25. G. Popescu, *Quantitative Phase Imaging of Cells and Tissues*, McGraw-Hill, New York (2011).
26. T. H. Nguyen and G. Popescu, "Spatial light interference microscopy (SLIM) using twisted-nematic liquid-crystal modulation," *Biomed. Opt. Express* **4**(9), 1571–1583 (2013).
27. Z. Wang et al., "Spatial light interference microscopy (SLIM)," *Opt. Express* **19**(2), 1016–1026 (2011).
28. B. Bhaduri et al., "Diffraction phase microscopy: principles and applications in materials and life sciences," *Adv. Opt. Photonics* **6**(1), 57–119 (2014).
29. G. Popescu et al., "Diffraction phase microscopy for quantifying cell structure and dynamics," *Opt. Lett.* **31**(6), 775–777 (2006).
30. E. Cuche, F. Bevilacqua, and C. Depeursinge, "Digital holography for quantitative phase-contrast imaging," *Opt. Lett.* **24**(5), 291–293 (1999).
31. P. Bon et al., "Quadriwave lateral shearing interferometry for quantitative phase microscopy of living cells," *Opt. Express* **17**(15), 13080 (2009).
32. J. W. Goodman, *Speckle Phenomena in Optics: Theory and Applications*, Roberts and Company Publishers, Greenwood Village (2007).
33. C. Edwards et al., "Epi-illumination diffraction phase microscopy with white light," *Opt. Lett.* **39**(21), 6162–6165 (2014).
34. Z. Wang et al., "Tissue refractive index as marker of disease," *J. Biomed. Opt.* **16**(11), 116017 (2011).
35. S. Sridharan et al., "Prediction of prostate cancer recurrence using quantitative phase imaging," *Sci. Rep.* **5**, 9976 (2015).
36. B. Bhaduri et al., "Cardiomyocyte imaging using real-time spatial light interference microscopy (SLIM)," *PLoS One* **8**(2), e56930 (2013).
37. B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature* **290**, 91–97 (1981).
38. T. Leung and J. Malik, "Recognizing surfaces using three-dimensional textons," in *Proc. of the Seventh IEEE Int. Conf. on Computer Vision* (1999).
39. T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vision* **43**(1), 29–44 (2001).
40. J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, Alaska (2008).
41. J. Shotton et al., "TextonBoost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vision* **81**(1), 2–23 (2009).
42. J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *J. Opt. Soc. Am. A* **7**(5), 923–932 (1990).
43. J. Malik et al., "Contour and texture analysis for image segmentation," *Int. J. Comput. Vision* **43**(1), 7–27 (2001).
44. L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).
45. L. Breiman, "Bagging predictors," *Mach. Learn.* **24**(2), 123–140 (1996).
46. J. Shotton et al., "Real-time human pose recognition in parts from single depth images," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'11)*, Colorado Springs (2011).
47. F. Schroff, A. Criminisi, and A. Zisserman, "Object class segmentation using random forests," in *Proc. of the British Machine Vision Conf.*, pp. 1–54, BMVA Press (2008).
48. B. Glocker et al., "Joint classification-regression forests for spatially structured multi-object segmentation," in *European Conf. on Computer Vision (ECCV'12)*, pp. 870–881, Springer, Berlin (2012).
49. M. Sun, P. Kohli, and J. Shotton, "Conditional regression forests for human pose estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 3394–3401 (2012).
50. A. Criminisi et al., "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Med. Image Anal.* **17**(8), 1293–1303 (2013).
51. D. Zikic et al., "Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI'12)*, pp. 369–376, Springer, Heidelberg (2012).
52. A. Criminisi and J. Shotton, *Decision Forests for Computer Vision and Medical Image Analysis*, Springer Science & Business Media (2013).
53. P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.* **63**(1), 3–42 (2006).
54. "MexOpenCV," http://vision.is.tohoku.ac.jp/kyamagu/software/mexopencv/
55. L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(6), 583–598 (1991).
56. W. C. Allsbrook et al., "Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist," *Hum. Pathol.* **32**(1), 81–88 (2001).
57. J. A. Tuxhorn et al., "Reactive stroma in human prostate cancer induction of myofibroblast phenotype and extracellular matrix remodeling," *Clin. Cancer Res.* **8**(9), 2912–2923 (2002).
58. D. A. Barron and D. R. Rowley, "The reactive stroma microenvironment and prostate cancer progression," *Endocr.-Relat. Cancer* **19**(6), R187–R204 (2012).
59. Z. Wang, H. Ding, and G. Popescu, "Scattering-phase theorem," *Opt. Lett.* **36**(7), 1215–1217 (2011).
60. M. B. Amin et al., *Gleason Grading of Prostate Cancer: A Contemporary Approach*, Lippincott Williams and Wilkins, Philadelphia (2003).
61. D. A. Freedman, *Statistical Models: Theory and Practice*, Cambridge University Press, New York (2009).

Tan H. Nguyen is currently a senior imaging engineer at Butterfly Network Inc. He obtained his BEng degree in electrical engineering from Ho Chi Minh City University of Technology, Vietnam, in 2009, his master's degree in signal processing, and his PhD in biomedical optics both from the University of Illinois of Urbana–Champaign (UIUC) in 2012 and 2016, respectively. His research interest includes biomedical imaging, inverse problems in optics, signal processing, and machine learning.

Shamira Sridharan received her PhD in bioengineering from the UIUC in 2015. Her graduate research focused on the application of quantitative phase imaging (QPI) for pathology to improve disease diagnosis and prognosis. Her research interests include clinical applications of optics and studying the effect of microenvironment on cellular proliferation.

Virgilia Macias is a research pathologist with background in oncologic surgical pathology, training in electron microscopy and immunohistochemistry. She did her training in Mexico. She is currently a research assistant professor in the Transdisciplinary Pathology Department at the University of Illinois at Chicago. Her research interest is mainly focused on investigation and/or validation of prostate cancer biomarkers. She is involved in laser microdissection, quantitative image analysis, tissue microarray (TMA) construction, and tissue banking.

Andre Kajdacsy-Balla is a professor and director of transdisciplinary pathology at the University of Illinois at Chicago. His clinical focus is anatomic pathology with special interest in gynecologic pathology and prostate cancer. His research interests are in the areas of tissue banking, TMAs, application of molecular techniques to tissue pathology, prostate cancer clinical outcomes prediction methods, and the effect of environmental agents on prostate cancer progression and metastasis.

Jonathan Melamed is a professor at New York University School of Medicine and director of anatomic pathology at one of its associated hospitals. His clinical responsibilities and expertise include service as a uropathologist. His research interests are in biobanking and the

assessment of prognostic and predictive biomarkers in urologic tumors with specific emphasis on prostate cancer.

Minh N. Do received his BEng degree in computer engineering from the University of Canberra, Australia in 1997 and his DrSci degree in communication systems from the Swiss Federal Institute of Technology, Lausanne, in 2001. Since 2002, he has been a professor in the Department of Electrical and Computer Engineering, UIUC and holds joint appointments with the Coordinated Science Laboratory, the Beckman Institute for Advanced Science and Technology, and the Department of Bioengineering.

Gabriel Popescu received his PhD in optics from the School of Optics/CREOL (now the College of Optics and Photonics), UCF in 2002. He continued with Michael Feld at Massachusetts Institute of Technology as a postdoctoral associate. In 2007, he joined ECE and the Beckman Institute at UIUC, where he has been active in QPI, on which he authored a book (McGraw-Hill, 2011). He founded Phi Optics, a start-up company that commercializes QPI technology. He is an OSA and SPIE fellow.