Abstract—Radical prostatectomy is performed on approximately 40% of men with organ-confined prostate cancer. Pathologic information obtained from the prostatectomy specimen provides important prognostic information and guides recommendations for adjuvant treatment. The current pathology protocol in most centers involves primarily qualitative assessment. In this paper, we describe and evaluate our system for automatic prostate cancer detection and grading on hematoxylin & eosin-stained tissue images. Our approach is intended to address the dual challenges of large data size and the need for high-level tissue information about the locations and grades of tumors. Our system uses two stages of AdaBoost-based classification. The first provides high-level tissue component labeling of a superpixel image partitioning. The second uses the tissue component labeling to provide a classification of cancer vs. non-cancer, and low-grade vs. high-grade cancer. We evaluated our system using 991 sub-images extracted from digital pathology images of 50 whole-mount tissue sections from 15 prostatectomy patients. We measured accuracies of 90% and 85% for the cancer vs. non-cancer and high-grade vs. low-grade classification tasks, respectively. This system represents a first step toward automated cancer quantification on prostate digital histopathology imaging, which could pave the way for more accurately informed post-prostatectomy patient care.

Index Terms—Automated prostate cancer detection, cancer grading, digital pathology image analysis, quantitative pathology, superpixels, machine learning

I. INTRODUCTION

PROSTATE CANCER (PCa) is the most common non-cutaneous cancer among men and radical prostatectomy, in which the entire prostate is surgically removed, is performed on approximately 40% of men with organ-confined PCa [1]. The post-prostatectomy pathologic assessment of the resected specimen by the pathologist yields crucial prognostic information that predicts surgical success and guides recommendations for adjuvant therapy [2]. Each tumor is assessed for its location within the prostate, volume, degree of differentiation (using the Gleason grading system [3]), extension into the seminal vesicles (termed as seminal vesicle invasion; SVI) or beyond the prostate (termed as extra-prostatic extension; EPE), and existence at the surgical resection margin (termed as a positive surgical margin; PSM). The prognosis of a patient is known to be related to the volumes and Gleason grades of the tumors observed in the resected specimen, as well as on SVI, EPE, and PSM status [2]. Adjuvant therapies such as radiation or hormone therapy may be considered for individuals with adverse pathology features.

Tumor volume and EPE status, in particular, are challenging to report quantitatively, and substantial inter-observer variability has been reported in the methods used clinically to make these assessments [4], [5]. The advent of high-resolution (e.g. 0.25 µm/pixel) whole-slide scanners is fostering a transition to a digital pathology workflow, similar to the transition from light-box viewing of film images to digital imaging in radiology. This transition opens the possibility for the integration of computational tools into the digital pathology workflow in order to enable quantitative assessments and reporting in a clinically feasible fashion. In the post-prostatectomy prostate cancer setting, the quantitative reporting of tumor grade, location, volume, SVI, EPE, and PSM status depends on an accurate classification of each local tissue region as being cancerous or benign.

Prostate cancer detection on H&E-stained prostate tissue images and assessment of cancer grade are challenging due to the complexity of appearance of normal tissue and cancerous tissue of different grades, as illustrated in Fig. 1(a-c). To perform this task accurately, experts employ high-level knowledge regarding the expected morphologic, geometric, and color attributes of normal and cancerous tissues. Another challenge to this assessment pertains to the sizes of the tissue structures that need to be examined, relative to the usual size of a human prostate. Pathologists frequently assess prostate tissue at a magnification corresponding to a pixel size of 0.5 µm/pixel in a scanned image. To illustrate a typical data size encountered, each prostate section can be approximately 4 cm × 3 cm in size, yielding an image of 80,000 × 60,000 pixels. With approximately 2–6 such images obtained from each prostate, tens of billions of RGB values are usually obtained in total. Thus, this classification problem poses the dual challenges of
the need for high-level tissue information and the processing of very large data sets.

To address these challenges, in this paper, we describe and quantitatively evaluate a software system for automatic prostate cancer detection on H&E-stained prostate tissue images. We preprocess all H&E-stained prostate tissue images by automatically partitioning each image into a set of non-overlapping superpixel regions using the method proposed in [6]. Pixels within each superpixel are similar in color and texture, and superpixel boundaries are aligned with intensity edges in the image. The superpixel partitioning provides a decrease in data size of several orders of magnitude while encoding potentially useful morphometric, geometric, and appearance information within the superpixels themselves. After superpixel partitioning, our system employs a two-stage classification. The first classification stage assigns to each superpixel a tissue component label (e.g. stroma, lumen, epithelial nuclei) having semantic meaning to an expert pathologist based on morphometric, geometric, and appearance information contained within and around each superpixel. Thus, we compute higher-level tissue information from the low-level image pixels. The second stage classifies image regions (each containing multiple labeled superpixels) as either low-grade cancer, high-grade cancer, or non-cancer based on local histograms of tissue component labels within each sub-region.

Specifically, we hypothesize that (1) an AdaBoost classifier trained on morphometric, geometric, and appearance attributes of superpixels will classify superpixels into tissue components with an accuracy of ≥ 80%; (2) an AdaBoost classifier trained on tissue component histograms of superpixel labels will classify 0.3 mm × 0.3 mm image sub-regions as cancer or non-cancer with an accuracy of ≥ 90% and a false negative rate of ≤ 15%; and (3) an AdaBoost classifier trained on tissue component histograms of superpixel labels will classify 0.3 mm × 0.3 mm cancerous image sub-regions as high-grade (Gleason 4) or low-grade (Gleason 3) cancer with an accuracy of ≥ 85% and a false negative rate of ≤ 10% for high-grade cancer detection.

II. RELATED WORK

There has been a substantial research focus on the problem of automatically detecting and grading prostate cancer on digital histopathology imaging; these research efforts have yielded valuable insights into the nature of this problem. The common approach is to compute feature information from the images, and then train a classifier to distinguish cancer from non-cancer, or to distinguish cancers of different grades, based on the computed features.

Within the context of the prior work, we distinguish between low-level and high-level features. Low-level features contain information about local distributions of pixel intensities, color and texture, intensity edges and their local orientations. Such features can be directly computed from images using standard image processing techniques, but have little or no semantic meaning to a pathologist. High-level features contain structural information about the content present in the image, such as tissue components (e.g. glands, lumina, stroma, nuclei), their
shape, color, size and geometric arrangement. These features have semantic meaning to a pathologist and are used in practice to guide the assessment [7]. In the following summary, we classify previously developed methods according to this distinction between low- and high-level features.

The majority of methods relying on low-level features utilize a multi-scale/multi-resolution approach and resort to machine learning techniques to train a classifier [8], [9], [10], [11], [12]. Doyle et al. [8] employed decision trees and trained a multi-scale Adaboost classifier to classify image pixels as either cancer or non-cancer based on 600 texture features. When tested on 22 images from 22 subjects, their method obtained 88% overall accuracy. Similarly, Doyle et al. [9] trained a multi-resolution Bayesian classifier with AdaBoost and tested on 100 biopsy cores from 58 subjects, obtaining 74% pixel accuracy in the cancer vs. non-cancer classification task. Diamond et al. [10] calculated morphological and texture characteristics on samples from 12 subjects to distinguish cancer from non-cancer images with 79% accuracy; they noted that long processing times were problematic. Huang et al. [11] found that multi-scale fractal dimension features were useful for prostate cancer grading, with 90%-94% classification accuracy; the number of subjects in their study, the origin of the tissue (biopsy or surgical specimen), and the false positive/negative rates were not specified. Khurd et al. [12] demonstrated that multi-scale texton characterization of the tissue images can be helpful to prostate cancer grading, obtaining 94% accuracy; the number of subjects in their study was not specified.

Several recent methods aimed to extract high-level structural information from the histopathology images [13], [14], [15]. Ideally, such high-level features should correspond to the information assessed by pathologists. In a paper focused on automated cancer grading, Doyle et al. [13] computed a set of graph-based features relying on manually labeled gland and nucleus centroid locations. Their automated grading experiment used 11 grade 3 and 7 grade 4 images and yielded an accuracy of 76.9% in distinguishing the two grades; the number of subjects in their study was not specified. This same research group reported in [14] improved results of 80% accuracy on a data set consisting of 16 grade 3 and 11 grade 4 images using a method intended to automate the selection of gland and nucleus centroids. The authors indicated that evaluation of their method on a larger data set is the subject of future work; the number of subjects in this study was not specified, and it is not clear if this was a superset of the data set used in [13]. Arif et al. [15] have also developed a method intended for automated nucleus extraction, with the ultimate aim of potentially supporting computer-aided diagnosis on prostate histopathology in the future. It is not clear how to generalize the methods in [14], [15] to other prostate tissue components (e.g. stroma, epithelium cytoplasm, intraluminal secretions, etc.), which are often used by pathologists for assessment.

Several other methods explored the utility of high-level features for automatic prostate histopathology assessment. Wittke et al. [16] manually segmented input images into epithelium, lumen, and stroma, and used derived features to distinguish low- from high-grade cancer; they reported 67% accuracy on images from 78 subjects. Xu et al. [17] used thresholding in the hue-saturation-intensity color space to segment the lumen and nuclei prior to support vector machine-based classification of high-grade vs. low-grade cancer, and reported 75% accuracy. Farjam et al. [18] demonstrated that explicit incorporation of knowledge specific to the domain of prostate histopathology assessment into a procedure for the segmentation of the images into stroma, lumen, and nuclei and subsequent computation of domain-specific features can yield more than 95% accurate classification of cancer grades. The number of subjects used was not specified for both of the above studies.

Most relevant to our work is the method of Tabesh et al. [19] developed by Aureon Laboratories Inc. The authors took a hybrid low-level/high-level approach and investigated the utility of color, texture, and morphometric features for distinguishing cancer from non-cancer and low-grade from high-grade cancer on tissue microarray cores. This method achieved 97% accuracy for the cancer vs. non-cancer classification task when tested on a set of 367 images and 81% accuracy for the high-grade vs. low-grade classification when tested on a set 218 images, respectively. The method was applied to relatively large $1600 \times 1200$ images, while some of the images were reported to have as little as 5% of the area covered by cancer, making direct tumor volume quantification difficult. Furthermore, the approach required approximately 30 minutes of processing time for each $1600 \times 1200$ image; pathologists wishing to use the (now closed) company’s analysis services were to physically ship biopsy samples for offline analysis.

Methods using low-level features along with multi-scale/multi-resolution approaches may result in the extraction of features which, although not directly meaningful to a pathologist, may correlate to high-level structures of pathological interest. The results of the described previous work, however, point to the idea that designing methods that directly extract high-level information from the image data may be beneficial to the task of automated prostate cancer detection and classification.

Although previous work has made important discoveries and strides toward a practically useful solution to this problem, a clinically-adopted solution remains elusive. Also, the lack of a standardized data set for this problem challenges the direct performance comparison of methods. For a system to be clinically adopted, it may be beneficial: (1) to classify tissue based on attributes that are used by and therefore intuitive to pathologists, improving their confidence in the software system; (2) to be robust and general via the avoidance of direct incorporation of potentially brittle domain-specific knowledge into the design of the system (as opposed to using expert labeling for training); and (3) to be evaluated on a data set with well-known subject characteristics and a fully described reference standard. Our proposed approach meets all three criteria. Without incorporating any domain-specific knowledge, our approach provides a fully automatic partitioning of the image into intermediate-size superpixel regions and assigns to each superpixel one of nine high-level tissue component labels having semantic meaning to pathologists. Our system for prostate cancer detection and grading based on these higher-
level labels was evaluated against a well-described reference standard data set, and with appropriate statistical inferences to account for the size of the data set and variability of performance of our system.

III. METHODS

Our approach consists of two main components. The first assigns each image pixel a high-level tissue component label based on a superpixel partitioning and low-level superpixel features. The second classifies image sub-regions as cancer or non-cancer, and as high-grade or low-grade cancer, based on high-level tissue component information. These two components are described in the following subsections.

A. Tissue Component Classification

The objective of tissue component classification is to assign to each image pixel a tissue component label having semantic meaning to an expert pathologist. We used nine tissue component labels: stroma, lumen, epithelial nucleus, epithelial cytoplasm, lymphocyte, red blood cell (RBC), intraluminal secretion, corpus amylaceum, and other. Figure 2 shows a general overview of the tissue component classification method. It is performed. The trained adaboost classifier (d) is then used to classify image superpixels and assign each superpixel a tissue component label (h).

method in [6] since it proposed a principled approach to compute superpixels with regular shapes and sizes in an energy minimization framework. Superpixels with regular shapes are less likely to straddle object (tissue component) boundaries, since such boundaries are mostly smooth. Moreover, by controlling the maximum size of each superpixel, we can influence the overall error in superpixel tessellation, where error is defined to be a superpixel split between two objects (tissue components).

The basic superpixel algorithm is illustrated in Figure 3. An image is covered with overlapping square patches of fixed size (Figure 3 left). Each pixel is contained in several patches, and we need to assign each pixel to one of the patches, thereby producing a superpixel tessellation. If two nearby pixels are assigned to the same patch, there is no penalty. If they belong to different patches, then there is a discontinuity penalty to pay. We set this penalty to be inversely proportional to the image gradient between these pixels. This encourages smoother, regularized boundaries that are well aligned with the intensity edges present in the image. The maximum superpixel size is equal to the patch size. Small superpixels are discouraged because they contribute a higher discontinuity penalty, since longer boundaries are required. Thus, the sizes of the superpixels are also regularized. A sample partitioning result is shown in Figure 3, right. This basic algorithm can be extended to other formulations, which allow a trade-off between a less regular spatial tessellation but more accurate boundaries or better efficiency. A complete description of this superpixel algorithm can be found in [6].

2) Superpixel Features: We computed a spatially dense set of scale invariant feature transform (SIFT) descriptors [24] for each of the images in our training set using the code provided by [25]. This resulted in a 128-feature vector \( v_p \) per image pixel \( p \in \Omega \). Similarly to [26], we used K-means, \( K = 100 \) to cluster all the descriptors from all the training images into a set of SIFT representatives \( \{V_k\}_{k=1}^K \). We labeled each image pixel \( p \in \Omega \) with the index \( i_p \) of the closest SIFT representative, namely

\[
i_p = \arg\min_{k \in \{1...K\}} ||v_p - V_k||_2.
\]

Next, we applied superpixel partitioning [6] to all the images in our training set. Each superpixel was then represented by a...
set of appearance, morphometry and geometry features, fully described in Table I. The appearance features describe the distribution of the color/intensity within each superpixel, including 5-bin per-channel intensity histograms. The morphometry features describe the shape, size and relative aspect ratio of each superpixel. Finally, the geometry features describe the distribution of the local gradients within superpixels. To that end, we used K-bin histograms over SIFT representative labels $i_p$ within each superpixel.

In addition to characterizing the appearance and shape of each superpixel, we also characterized the appearance of the neighborhood around each superpixel. Let $S \subseteq \Omega$ be a superpixel. Let $p_S$ be the centroid of the superpixel and $r$ be a radius in pixels, $r \in \{10, 20, 30\}$. We defined three rings of neighborhood

$$N_r(p_S) = \{ p \in \Omega | r - 10 \leq ||p - p_S|| \leq r, p \notin S \}.$$  

We then computed color and SIFT histograms (the last two rows in Table I) within these neighborhood rings and appended those histograms to the feature vector of each superpixel. Figure 4 shows an illustration of the neighborhoods $N_r$ for a superpixel and the construction of its feature vector.

3) Using AdaBoost for classifying superpixels: We used an Adaptive Boosting (AdaBoost) machine learning framework [27] to learn the superpixel descriptors characterizing each tissue component type based on the features described above. AdaBoost is a meta-algorithm; it can be used in conjunction with many different weak classifiers to improve their performance. Specifically, we use Real AdaBoost [20] that utilizes simple decision trees as weak learners. We trained nine binary classifiers, one per tissue component type. Each component-specific binary classifier assigns each superpixel either the “class” or “non-class” label for that component type (e.g. “stroma” or “not stroma”), along with a confidence value. The outcome of the nine binary classifiers was combined into one multi-label classifier by choosing the component label with the highest confidence.

B. Prostate Cancer Detection and Classification

Below, we describe our method for prostate tissue classification. This method relies on the tissue component classification described in Section III-A as a pre-processing step. We assume that for each image, a superpixel partition is computed, features are extracted and each superpixel is classified as a specific tissue component type. We then counted the number of pixels in the image labeled with each tissue component and compute a nine-bin tissue component histogram per image. We used tissue component histograms as feature vectors for cancer detection. Namely, we considered the task of binary “cancer” vs. “non-cancer” classification using boosted decision trees (Modest AdaBoost) [20].

We further focused on cancer tissue classification problem and considered the task of binary “high-grade” vs. “low-grade” classification, again using tissue component histograms as feature vectors and Modest AdaBoost [20].

IV. EXPERIMENTS

A. Materials

The image data for our experiments was acquired as part of a study that was approved by the research ethics board of our institution; written informed consent was obtained from all subjects. All subjects were suitable for and consented to radical prostatectomy, and had histologically confirmed prostate cancer (clinical stage T1 or T2). For each of 15 subjects, following radical prostatectomy, the resected prostate was fixed in 10% buffered formalin for 48 hours. Each specimen was then transversely sliced into 4.4 mm-thick sections. The
sections were processed using standard paraffin embedding, yielding whole-mount H&E-stained microscope slides, each containing a single 4 µm-thick section of tissue taken from each paraffin block face. The slides were digitized using a ScanScope GL (Aperio Technologies, Vista, CA, USA) bright field slide scanner. The acquired images were 24-bit color with isotropic 0.5 µm pixels. From each subject, between 2 and 5 (median 3) whole-mount sections were obtained; 50 such sections were obtained in total.

A physician (M.G.) trained by two genitourinary pathologists (M.M. and J.A.G.) assessed each image using the ScanScope ImageScope v11.0.2.725 software (Aperio Technologies, Vista, CA, USA) and contoured every tumor focus, as well as any identified areas of benign prostatic hyperplasia (BPH), atrophy (ATR), and prostatic intraepithelial neoplasia (PIN) using a Cintiq 12WX pen-enabled display (Wacom Co. Ltd., Saitama, Japan). Tumor areas were classified as either Gleason grade 3 (G3), 3+4 (G3+4), 4+3 (G4+3) or 4 (G4). G3+4 referred to a region that contained a mixture of Gleason grade 3 and Gleason grade 4 cancer, with more than 50% of the region comprising Gleason grade 3. G4+3 referred to a region that contains a mixture of Gleason grade 3 and Gleason grade 4 cancer, with more than 50% of the region comprising Gleason grade 4. No Gleason grade 5 was observed in our data set. An illustrative example of this contouring is provided in Figure 1. Contouring was performed at the highest image resolution, with the objective of generating contours enclosing regions consisting purely of the designated label (e.g. a “G4” contour is intended to contain only tissue that is cancerous with Gleason grade 4, devoid of normal tissue or cancer of other grades). Contouring and assessment in this fashion required approximately 70 hours of operator time per subject; Figure 1(b) and (c) provide an illustration of the attention to detail applied to this contouring task. All contours and assessments were performed by the trained physician (M.G.) and verified by a genitourinary pathologist (M.M. or J.A.G.).

From these images, we extracted 991 0.3 mm × 0.3 mm sub-images sampled onto a 301 pixel × 301 pixel grid (i.e. the pixel size was approximately 1 µm × 1 µm). Each sub-image resided entirely within one of the contoured regions of interest and thus inherited a corresponding label. Table II summarizes the distribution of sub-image labels in this set. This data set was used for the prostate cancer detection and classification experiment described in Section IV-C.

From the data set described in Table II, we extracted a subset described in Table III for use in the tissue component classification experiment described in Section IV-B. For each of these images, we performed a superpixel partitioning as described in Section III-A1. A custom user interface developed in-house was used by a physician (M.G.) trained by two genitourinary pathologists (M.M. and J.A.G.) to label a portion of superpixels in each image in the set. Each labeled superpixel was given one of the following nine tissue component labels: stroma, lumen, epithelial nucleus, epithelial cytoplasm, lymphocyte, red blood cell, secretion, corpus amylaceum, other. For superpixels containing tissue falling into more than one of the preceding categories, the physician applied the label representing the majority of the superpixel’s contents. A total of 63,926 superpixels were manually labeled in this fashion. A sample of such a manual labeling is provided in Figure 1(d).

### B. Tissue Component Classification

Because classification of tissue into components (stroma, lumen, epithelial nucleus, epithelial cytoplasm, lymphocyte, red blood cell, secretion, corpus amylaceum, other) is an integral part of our overall method for prostate cancer detection and classification, our first experiment evaluates tissue component classification in isolation from the other components of our system. We performed ten repeated subsampling cross-validation trials using the 63,926 manually-labeled superpixels from the data set described in Table III. In each trial, we randomly split the set of labeled superpixels into a training set comprising 80% of the labeled superpixels, and a test set comprising the remaining 20%. We performed a superpixel partitioning as described in Section III-A1. For all superpixels, we calculated superpixel features as described in Section III-A2. We then performed superpixel classification using AdaBoost as described in Section III-A3. AdaBoost requires two parameters: the maximal depth $d$ for tree learners and the maximal number $I$ of iterations. In this experiment, we used the default $d = 3$ and $I = 179$ (the number of features per superpixel). For each of the ten trials, nine binary AdaBoost classifiers (one per each tissue component type) were trained and tested, resulting in nine confidence values (one per label) for each superpixel in the test set. Each superpixel was assigned the label having the largest confidence value.

We measured the mean ± standard deviation of the multiclass classification error (the number of incorrectly labeled superpixels in the test set, divided by the total number of superpixels in the test set) across the ten trials as well as the confusion matrix for each trial. As a qualitative assessment of the tissue component labeling, we calculated a set of histograms showing the distribution of tissue components (stroma, lumen, epithelial nucleus, epithelial cytoplasm, lymphocyte, red blood cell, secretion, corpus amylaceum, other) within each prostate tissue type (Atrophy, PIN, BPH, G3, G3+4, G4+3, G4). This was done in order to evaluate the agreement of these observed distributions against pathologists’ knowledge of the expected distributions. As a further qualitative assessment of the tissue component labeling, we also rendered a set of images with classifier outputs color-coded, to compare the resulting superpixel labeling to the visible tissue components in the H&E-stained test images. Finally, to test hypothesis (1) (Section I), that our method will classify superpixels into tissue

---

**Table II**

<table>
<thead>
<tr>
<th>Tissue type</th>
<th>Number of sub-images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrophy (ATR)</td>
<td>493</td>
</tr>
<tr>
<td>Benign prostatic hyperplasia (BPH)</td>
<td>104</td>
</tr>
<tr>
<td>Prostatic intraepithelial neoplasia (PIN)</td>
<td>216</td>
</tr>
<tr>
<td>Gleason 3 (G3)</td>
<td>99</td>
</tr>
<tr>
<td>Mixed Gleason 3 and 4, majority 3 (G3+4)</td>
<td>46</td>
</tr>
<tr>
<td>Mixed Gleason 3 and 4, majority 4 (G4+3)</td>
<td>12</td>
</tr>
<tr>
<td>Gleason 4 (G4)</td>
<td>21</td>
</tr>
</tbody>
</table>
TABLE III 
NUMBER OF SUB-IMAGES PER PROSTATE TISSUE TYPE IN THE DATA SET USED FOR TISSUE COMPONENT CLASSIFICATION EXPERIMENT

<table>
<thead>
<tr>
<th>Tissue type</th>
<th>Number of sub-images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrophy</td>
<td>21</td>
</tr>
<tr>
<td>Benign prostatic hyperplasia (BPH)</td>
<td>17</td>
</tr>
<tr>
<td>Prostatic intraepithelial neoplasia (PIN)</td>
<td>17</td>
</tr>
<tr>
<td>Gleason 3 (G3)</td>
<td>20</td>
</tr>
<tr>
<td>Mixed Gleason 3 and 4, majority 3 (G3+4)</td>
<td>12</td>
</tr>
<tr>
<td>Mixed Gleason 3 and 4, majority 4 (G4+3)</td>
<td>5</td>
</tr>
<tr>
<td>Gleason 4 (G4)</td>
<td>5</td>
</tr>
</tbody>
</table>

components with accuracy of ≥ 80%, we performed a one-tailed t-test of the null hypothesis that the mean multiclass classification error ≥ 0.2.

C. Prostate Cancer Detection and Classification

This section describes two experiments. The first is on prostate cancer detection; the classification of sub-images as containing cancerous tissue, or non-cancerous tissue (Section IV-C1). The second is on prostate cancer classification; the classification of cancer-containing sub-images as containing high-grade cancer, or low-grade cancer (Section IV-C2). Both experiments were conducted using the 991-sub-image data set described in Table II. A superpixel partitioning was computed for each sub-image as described in Section III-A1. The number of superpixels per sub-image was recorded in order to measure the reduction in data size achieved by this partitioning. Each superpixel was then represented by a set of features as described in Section III-A2, and one of the nine tissue component labels was applied to each superpixel as described in Section III-A3. A tissue component histogram was computed for each sub-image and used for training and testing of a Modest AdaBoost classifier, as described in Section III-B. To evaluate classifier performance, we computed the false positive (FP), false negative (FN), true positive (TP), true negative (TN) and accuracy (TP + TN) rates.

1) Prostate Cancer Detection: This experiment evaluated the ability of a Modest AdaBoost classifier to classify sub-images as containing cancer or non-cancer, based on tissue component histograms. We performed ten repeated subsampling cross-validation trials using the 991 sub-images in the data set described in Table II. In each trial, we randomly split the set of labeled superpixels into a training set comprising 80% (792) of the sub-images, and a test set comprising the remaining 20% (199). For this experiment, each sub-image in the data set was labeled as cancer (“positive”) or non-cancer (“negative”). Any sub-image having a label of Atrophy, BPH, or PIN was designated as non-cancer. Any sub-image having a label of G3, G3+4, G4+3, or G4 was designated as cancer.

Since, in the clinical workflow, a failure to detect cancer may result in the denial of necessary post-prostatectomy treatment, it may be beneficial to tune the classifier to achieve a decreased FN rate, even if compromising in terms of an increased FP rate. Depending on the classifier at hand, this can be done by either skewing the error cost matrix, or by using unbalanced class probability priors, or both. In these experiments, due to the relatively lower number of positive data points, we implicitly skewed the error cost matrix by artificially augmenting the training set with an additional number s of duplicates for each positive sub-image in the set. This way, during classifier training, each FN error was counted s times compared to FP errors. Thus, for s > 0, FN errors carried heavier weight with the intention of influencing the classifier training toward obtaining a lower FN rate. We varied the number of duplicates used in training in order to observe how the classification results changed as a function of number of duplicates. To this end, each of the cross-validation experiments was performed five times, one for each s ∈ {0, 5, 10, 15, 20}.

We performed the following measurements for all values of s. We measured the mean ± standard deviation (where applicable) of the classification error across all cross-validation trials. Classification error was calculated as the number of incorrectly labeled sub-images in the test set, divided by the total number of sub-images in the test set. In addition to the FN and FP rates, we measured the TP and TN rates. We also plotted a precision-recall curve showing the mean TN vs. mean TP (averaged across all trials) as a function of s in order to visualize the trade-off resulting from increasing the number of positive duplicates in the data set. To test hypothesis (2) (Section I), that our method will classify sub-images as cancer or non-cancer with accuracy of ≥ 90% and a false negative rate of ≤ 15%, we performed one-tailed t-tests of the null hypotheses that the mean classification error ≥ 0.1 and that the FN rate ≥ 0.15.

2) Prostate Cancer Classification: This experiment evaluated the ability of a Modest AdaBoost classifier to classify cancer-containing sub-images as containing high-grade cancer or low-grade cancer, based on tissue component histograms. It was conducted identically to the prostate cancer detection experiment described in Section IV-C1, with the exception of the following details. We performed a leave-one-out cross-validation experiment using the 120 G3 and G4 sub-images in the data set described in Table II. Leave-one-out cross-validation was performed due to the relatively small size of the data set for this experiment. For this experiment, each sub-image in the data set was labeled as low-grade cancer or high-grade cancer. Sub-images having a label of G3 were designated as low-grade cancer. Sub-images having a label of G4 were designated as high-grade cancer. In this experiment, for purposes of computing TP, TN, FP, and FN, high-grade cancer was considered as “positive” and low-grade cancer as “negative”. To test hypothesis (3) (Section I), that our method will classify sub-images as high-grade or low-grade cancer with accuracy of ≥ 85% and a false negative rate of ≤ 10% for high-grade cancer detection, we performed one-tailed t-tests of the null hypotheses that the mean classification error ≥ 0.15 and that the FN rate ≥ 0.1.

V. RESULTS

A. Tissue Component Classification

This section reports the results of the tissue component classification experiment described in Section IV-B. The mean ± standard deviation of the multiclass error rate over 10 trials
was 0.16 ± 0.009. The data passed a one-sample Kolmogorov-Smirnov test of normality ($p = 0.89$, $\alpha = 0.05$). A one-tailed t-test of the null hypothesis that the mean multiclass classification error $\geq 0.2$ yielded a p-value of $1.4 \times 10^{-10}$. As this test involved ten samples generated via cross-validation, a Bonferroni-corrected $\alpha$ of $0.05/10 = 0.005$ was used. The null hypothesis was therefore rejected. The Bonferroni-adjusted 95% confidence interval on the mean classification error was found to be (0.1598, 0.1687).

Figure 5 shows a classification confusion matrix for one of the cross-validation trials. The rows represent the actual tissue component labels and the columns represent the labels predicted by the Real AdaBoost classifier (Section IV-B). A cell $(i, j)$ represents the frequency of an event whereby a superpixel with the actual label $i$ was classified as label $j$. The rows of the matrix are normalized separately for each label (row) to sum to one. The numbers on the left of the figure represent the frequency of each label in the test set.

![Figure 5. Confusion matrix for the tissue component classification: Actual labels are the rows, predicted labels are the columns. A cell $(i, j)$ represents the frequency of an event whereby a superpixel with the actual label $i$ was classified as label $j$. The rows of the matrix are normalized separately for each label (row) to sum to one. The numbers on the left of the figure represent the frequency of each label in the test set.](image)

![Figure 6. Top: Distribution of tissue component labels per prostate tissue type. Bottom: The histograms are reorganized and for each tissue component we show how its relative proportion varies across different prostate tissue types.](image)
parallelizable. The mean ± standard deviation number of superpixels per sub-image for the data set described in Table II was $1396 ± 221$.

1) Prostate Cancer Detection: This section reports the results of the prostate cancer detection experiment described in Section IV-C1. The mean ± standard deviation of the error rate (TP+TN) over ten trials was $0.07 ± 0.01, 0.08 ± 0.02, 0.09 ± 0.02, 0.10 ± 0.02, 0.11 ± 0.02$, corresponding to $s ∈ \{0, 5, 10, 15, 20\}$, respectively. The data (aggregated over all $s ∈ \{0, 5, 10, 15, 20\}$) passed a one-sample Kolmogorov-Smirnoff test of normality ($p = 0.71, \alpha = 0.05$). A one-tailed t-test of the null hypothesis that the mean classification error $\geq 0.1$ yielded a p-value of $0.0003$. As this test involved 50 samples generated via ten cross-validation trials for each value of $s$, a Bonferroni-corrected $\alpha$ of $0.05/50 = 0.001$ was used. The null hypothesis was therefore rejected. The Bonferroni-adjusted 95% confidence interval on the mean classification error was found to be $(0.0780, 0.0994)$.

The mean ± standard deviation of the FN rate over ten trials was $0.22 ± 0.08, 0.06 ± 0.03, 0.05 ± 0.04, 0.04 ± 0.03, 0.03 ± 0.03$, corresponding to $s ∈ \{0, 5, 10, 15, 20\}$, respectively. The data (aggregated over all $s ∈ \{0, 5, 10, 15, 20\}$) passed a one-sample Kolmogorov-Smirnoff test of normality ($p = 0.07, \alpha = 0.05$). A one-tailed t-test of the null hypothesis that the mean false negative rate $≥ 0.15$ yielded a p-value of $< 0.0001$. As this test involved 50 samples generated via ten cross-validation trials for each value of $s$, a Bonferroni-corrected $\alpha$ of $0.05/50 = 0.001$ was used. The null hypothesis was therefore rejected. The Bonferroni-adjusted 95% confidence interval on the mean FN rate was found to be $(0.04, 0.12)$.

Figures 8-10 show the mean ± standard deviation over ten trials of FN, FP and accuracy (TP+TN) rates, respectively, as a function of the number $s$ of duplicates used for each positive (cancer) example in the training. Figure 11 shows the recall-precision curve.

Table IV shows the distribution of classification errors across prostate tissue types as a function of the number $s$ of duplicates used in the training data set. For each experiment corresponding to $s ∈ \{0, 5, 10, 15, 20\}$ and prostate tissue type, we show the proportion of the images with this type classified as cancer (C) or non-cancer (NC), averaged over ten cross-validation trials. As the number $s$ of positive (cancer) duplicates used in training increases, the number of cancer images misclassified as non-cancer decreases. The higher the grade of cancer, the faster the FN rate decreases. (Standard deviations can be seen in Figures 8-10 and are omitted from Table IV for clarity.)

2) Prostate cancer classification: This section reports the results of the prostate cancer classification experiment described in Section IV-C2. The cross-validation error rate was $0.08, 0.08, 0.09, 0.13, 0.13$, corresponding to $s ∈ \{0, 5, 10, 15, 20\}$, respectively. The data (aggregated over all $s ∈ \{0, 5, 10, 15, 20\}$) passed a one-sample Kolmogorov-Smirnoff test of normality ($p = 0.83, \alpha = 0.05$). A one-tailed t-test of the null hypothesis that the mean classification error $\geq 0.15$ yielded a p-value of $0.005$. As this test involved five samples generated via one leave-one-out cross-validation for
each values of \( s \), a Bonferroni-corrected \( \alpha \) of 0.05/5 = 0.01 was used. The null hypothesis was therefore rejected. The Bonferroni-adjusted 95% confidence interval on the mean classification error was found to be (0.0514, 0.1486).

The cross-validation FN rate was 0.05, 0.04, 0.04, 0.05, 0.04, corresponding to \( s \in \{0, 5, 10, 15, 20\} \), respectively. The data (aggregated over all \( s \in \{0, 5, 10, 15, 20\} \)) passed a one-sample Kolmogorov-Smirnov test of normality \((p = 0.41, \alpha = 0.05)\). A one-tailed t-test of the null hypothesis that the mean FN rate \( \geq 0.1 \) yielded a p-value of \(< 0.0001\). As this test

---

Fig. 8. Cancer detection average false negative rate: \( \text{FN}/(\text{FN}+\text{TP}) \). We show the average rate of non-cancer images classified as non-cancer, as a function of the number \( s \) of duplicates used for each cancer image in training. For each experiment, ten cross-validations were performed and the mean \( \pm \) standard deviation error rate is reported.

Fig. 9. Cancer detection average false positive rate: \( \text{FP}/(\text{FP}+\text{TN}) \). We show the average rate of non-cancer images classified as cancer, as a function of the number \( s \) of duplicates used for each cancer image in training. For each experiment, ten cross-validations were performed and the mean \( \pm \) standard deviation error rate is reported.

Fig. 10. Cancer detection average accuracy rate: \( \text{(TP+TN)}/(\text{FN+TP+FP+TN}) \). We show the average rate of correctly classified images, as a function of the number \( s \) of duplicates used for each cancer image in training. For each experiment, ten cross-validations were performed and the mean \( \pm \) standard deviation accuracy is reported.

Fig. 11. Cancer detection recall-precision rate curve: average \( \text{TN} \) vs. average \( \text{TP} \) as a function of the number \( s \) of duplicates used for each cancer image in training.

TABLE IV

Table: Average confusion matrix as a function of number of duplicates used in training. For each experiment and prostate tissue type, we show the proportion of the images with this type classified as cancer (C) or non-cancer (NC), averaged over ten trials. Numbers in bold show the proportion of images classified correctly.

<table>
<thead>
<tr>
<th>Number of duplicates ( s )</th>
<th>( s = 0 )</th>
<th>( s = 5 )</th>
<th>( s = 10 )</th>
<th>( s = 15 )</th>
<th>( s = 20 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATR</td>
<td>( C ) NC</td>
<td>( C ) NC</td>
<td>( C ) NC</td>
<td>( C ) NC</td>
<td>( C ) NC</td>
</tr>
<tr>
<td>BPH</td>
<td>0.00</td>
<td>1.00</td>
<td>0.02</td>
<td>0.98</td>
<td>0.03</td>
</tr>
<tr>
<td>PIN</td>
<td>0.20</td>
<td>0.80</td>
<td>0.42</td>
<td>0.58</td>
<td>0.48</td>
</tr>
<tr>
<td>G3</td>
<td>0.75</td>
<td>0.25</td>
<td>0.95</td>
<td>0.05</td>
<td>0.96</td>
</tr>
<tr>
<td>G3+4</td>
<td>0.71</td>
<td>0.29</td>
<td>0.87</td>
<td>0.13</td>
<td>0.87</td>
</tr>
<tr>
<td>G4</td>
<td>0.93</td>
<td>0.07</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>G4+3</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
involved five samples generated via one leave-one-out cross-validation for each value of \( s \), a Bonferroni-corrected \( \alpha \) of 0.05/5 = 0.01 was used. The null hypothesis was therefore rejected. The Bonferroni-adjusted 95% confidence interval on the mean FN rate was found to be (0.04, 0.05).

Figures 12-14 show the FN, FP and accuracy (TP+TN) rates, respectively, as a function of the number of duplicates used for each positive (high grade) image in the training computed with leave-one-out cross-validation. Figure 15 shows the recall-precision curve.

VI. DISCUSSION

A. Tissue Component Classification

Many computer vision applications (e.g. [28], [29], [30], [31], [32], [33], [34], [35]) have benefited from representing an image as a collection of superpixels. A superpixel is considered to be a perceptually meaningful, atomic, arbitrarily-shaped image subregion whose borders are intended to be better aligned with intensity edges than those of a rectangular region. Usually, a superpixel contains pixels that are similar in color and texture. Such pixels are likely to belong to the
same physical world object or, in our case, the same tissue component. Superpixel partitioning of an image reduces data dimensionality and can naturally be used for computing features that need spatial support [6]. To the best of the authors’ knowledge, this work represents the first application of a superpixel partitioning to the problem of prostate pathology image classification. In this work, the superpixel partitioning provided more than an order of magnitude reduction in the number of elements to be processed within each sub-image, from $301 \times 301 = 90,601$ pixels to a mean of 1396 superpixels per sub-image ($> 60$-fold reduction in data size). Morphometric, geometric, and appearance features computed based on the superpixel partitioning provided our tested classifier with the necessary data to accurately assign to each superpixel a higher-level tissue component label having semantic meaning to an expert pathologist. This result points to the suitability of the superpixel partitioning to this problem domain.

For classification, we chose to use AdaBoost as it is commonly used in computer vision, data mining, pattern recognition and medical imaging applications (e.g., [36], [37], [38], [39], [40], [41]) for its relative simplicity and speed. We observed that this classifier provided useful accuracy in tissue component labeling of superpixels, with the exception of an observed confounding effect of the “other” label (Figure 5). Further inspection revealed that the “other” label accounted for several additional, infrequently occurring, tissue subtypes of heterogeneous appearance. For example, it includes superpixels lying in the empty space surrounding glands, covering cristalloids or blood vessel walls. The main reason for the relatively high error rate for the “other” label is the low frequency of these sub-types in the training and test data, high variability of appearance and the lack of higher level information that allows for distinguishing e.g. the space surrounding gland from lumen (both uniformly white), or vessel walls from stroma (both red). We observed (Figure 6) that the proportions of tissue component labels for each prostate tissue type correspond to expected distributions based on expert knowledge. For example, it is known that the ratio of stroma to epithelial nuclei decreases with increasing Gleason grade, and this is reflected in Figure 6. The decrease in lumen with increasing Gleason grade shown in Figure 6 is also anticipated as a consequence of more advanced cancer.

Our one-tailed t-test of the null hypothesis that the mean multiclass classification error $\geq 0.2$ yielded a p-value sufficient for rejection of the null hypothesis. Thus, we assert that hypothesis (1) (Section I) is true. Our observed 95% confidence interval suggests that a mean classification error of not more than $\sim 0.16$, corresponding to an accuracy of $\sim 84\%$, can be expected from the tested method.

B. Prostate Cancer Detection and Classification

The results of our prostate cancer detection and classification experiments can be interpreted in the context of our reference standard data set, which contained meticulously drawn manual contours (e.g., Figure 1(a-c)) verified by a minimum of two experts. The quality of this data set contributed to the effectiveness of classifier training and validity of testing.

1) Prostate Cancer Detection: In our cancer detection experiment, we noted a general improvement in FN rate with increasing $s$, without a deleterious concomitant increase in FP rate (Table IV). However, this is not true for PIN; with increasing $s$, the accuracy for PIN images decreases rapidly with a high FP rate for PIN images. In fact, at $s = 20$, PIN images are classified as cancer 55% of the time and as non-cancer 45% of the time. Upon reflection, this result is unsurprising on account of the fact that PIN is widely considered to be tissue in a pre-cancerous state; not quite cancer, but certainly abnormal and with a predisposition to become cancerous. This leads to the observation that PIN images may not in fact be correctly classifiable as either cancer or non-cancer, and our overall FP rate for cancer may be improved by placing PIN within a distinct third category.

Our one-tailed t-test of the null hypothesis that the mean classification error $\geq 0.1$ yielded a p-value sufficient for rejection of the null hypothesis. Our one-tailed t-test of the null hypothesis that the FN rate $\geq 0.15$ yielded a p-value sufficient for rejection of the null hypothesis. Thus, we assert that hypothesis (2) (Section I) is true. Our observed 95% confidence interval suggests that a mean classification error of not more than $\sim 0.1$, corresponding to an accuracy of $\sim 90\%$, can be expected from the tested method. Our observed 95% confidence interval suggests that a mean FN rate of not more than $\sim 12\%$ can be expected from the tested method. As these confidence intervals were calculated including all tested numbers $s$ of duplicates in the training set, this expected classification error and FN rate should be robust with respect to the choice of $s$.

2) Prostate Cancer Classification: The results of our cancer classification experiment should be interpreted in the context of the relatively small number of images in the data set (99 low-grade and 21 high-grade). We restricted the data set for this experiment to only the sub-images having pure G3 or G4 tissue type, excluding the G3+4 and G4+3 sub-images. Our rationale for this exclusion was that since the contoured G3+4 and G4+3 regions contain a mixture of low and high-grade cancer, and since our sub-images are relatively small at $0.3 \times 0.3$ mm in size, using the G3+4 and G4+3 sub-images for high-grade vs. low-grade classification could introduce substantial pollution into the training and test sets. In this case, the reference standard labels for the images could be incorrect, challenging the unambiguous interpretation of the results. This issue does not disrupt the prostate cancer detection experiment, since G3+4 and G4+3 sub-images uniformly contain cancer.

We also observed oscillation in the precision recall curve (Figure 15) for this experiment (as compared with the precision recall curve for the prostate cancer detection experiment shown in Figure 11). This effect is possibly attributable to the relatively small size of our data set. Nevertheless, we did observe a reduction in the FN rate for high-grade cancer without a material concomitant rise in FP rate at $s = 5$ (Figures 12 and 13). Testing on a larger sample size would provide further evidence against which to compare these results. However, with the advent of prostate-specific antigen testing and 18-gauge needle core biopsy, prostate cancer is being diagnosed earlier, making the presence of high-grade cancer
at prostatectomy increasingly rare, challenging the acquisition of a large data set for high-grade vs. low-grade classification evaluation.

Our one-tailed t-test of the null hypothesis that the mean classification error $\geq 0.15$ yielded a p-value sufficient for rejection of the null hypothesis. Our one-tailed t-test of the null hypothesis that the FN rate $\geq 0.1$ yielded a p-value sufficient for rejection of the null hypothesis. Thus, we assert that hypothesis (3) (Section I) is true. Our observed 95% confidence interval suggests that a mean classification error of not more than $\sim 0.15$, corresponding to an accuracy of $\sim 85\%$, can be expected from the tested method. Our observed 95% confidence interval suggests that a mean FN rate of not more than $\sim 5\%$ can be expected from the tested method. As these confidence intervals were calculated including all tested numbers $s$ of duplicates in the training set, this expected classification error and FN rate should be robust with respect to the choice of $s$.

VII. CONCLUSION

We have designed and evaluated a software system for prostate cancer detection and classification on digitized, hematoxylin & eosin-stained digital histopathology images. Our system uses two stages of AdaBoost-based classification. The first provides high-level tissue component labeling of a superpixel partitioning of the images. The second uses the tissue component labeling to provide a classification of cancer vs. non-cancer, and low-grade vs. high-grade cancer. The superpixel partitioning provided a more than 60-fold reduction in data size, increasing processing efficiency. Using our database of 991 sub-images, our statistical testing measured accuracies of 90% and 85% for the cancer vs. non-cancer and high-grade vs. low-grade classification tasks, respectively. We also measured a false-negative (FN) rate for cancer detection of 12% and for high-grade cancer detection (in our high-grade vs. low-grade classification experiment) of 5%.

Our system determines the high-level tissue component labeling of superpixels without the use of any explicitly encoded domain knowledge, automatically learning the labeling from a training set. It therefore can potentially be trained to classify tissue components used by pathologists in analysis of other organs (e.g. breast, brain, etc.). To the best of the authors’ knowledge, this work represents the first application of a superpixel image partitioning to the problem of digital histopathology image processing. Tissue component labels were applied based on morphometric, geometric, and appearance information derived from the superpixel partitioning. These labels encode high-level information, similar to that used by pathologists for the task of cancer detection and classification, and were found to support robust automation of these tasks in this study. This system represents a first step toward automated cancer quantification on prostate digital histopathology imaging, which could pave the way for better informed post-prostatectomy patient care.

ACKNOWLEDGMENT

This work was supported by the Canadian Institutes of Health Research (funding reference number CTP 87515) and the Natural Sciences and Engineering Research Council of Canada. A. Fenster holds a Canada Research Chair (Tier I). A. D. Ward holds a Cancer Care Ontario Research Chair in Cancer Imaging.

REFERENCES


