J Pathol Inform

Editor-in-Chief:
Anil V. Parwani , Columbus, OH, USA
Liron Pantanowitz, Pittsburgh, PA, USA

OPEN ACCESS
HTML format

For entire Editorial Board visit : www.jpathinformatics.org/editorialboard.asp

Research Article

Empirical comparison of color normalization methods for epithelial-stromal classification in H and E images

Amit Sethi[1,2], Lingdao Sha[3], Abhishek Ramnath Vahadane[1], Ryan J. Deaton[2], Neeraj Kumar[1], Virgilia Macias[2], Peter H. Gann[2]

[1]Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam, India, Departments of [2]Pathology and [3]Electrical and Computer Engineering, University of Illinois, Chicago, IL, USA

E-mail: *Dr. Amit Sethi - amitsethi@iitg.ac.in
*Corresponding author

Received: 01 October 2015 Accepted: 04 February 2016 Published: 11 April 2016

Abstract

Context: Color normalization techniques for histology have not been empirically tested for their utility for computational pathology pipelines. **Aims:** We compared two contemporary techniques for achieving a common intermediate goal – epithelial-stromal classification. **Settings and Design:** Expert-annotated regions of epithelium and stroma were treated as ground truth for comparing classifiers on original and color-normalized images. **Materials and Methods:** Epithelial and stromal regions were annotated on thirty diverse-appearing H and E stained prostate cancer tissue microarray cores. Corresponding sets of thirty images each were generated using the two color normalization techniques. Color metrics were compared for original and color-normalized images. Separate epithelial-stromal classifiers were trained and compared on test images. Main analyses were conducted using a multiresolution segmentation (MRS) approach; comparative analyses using two other classification approaches (convolutional neural network [CNN], *Wndchrm*) were also performed. **Statistical Analysis:** For the main MRS method, which relied on classification of super-pixels, the number of variables used was reduced using backward elimination without compromising accuracy, and test - area under the curves (AUCs) were compared for original and normalized images. For CNN and *Wndchrm*, pixel classification test-AUCs were compared. **Results:** Khan method reduced color saturation while Vahadane reduced hue variance. Super-pixel-level test-AUC for MRS was 0.010–0.025 (95% confidence interval limits ± 0.004) higher for the two normalized image sets compared to the original in the 10–80 variable range. Improvement in pixel classification accuracy was also observed for CNN and *Wndchrm* for color-normalized images. **Conclusions:** Color normalization can give a small incremental benefit when a super-pixel-based classification method is used with features that perform implicit color normalization while the gain is higher for patch-based classification methods for classifying epithelium versus stroma.

Key words: Color normalization, computational pathology, epithelial-stromal classification

Access this article online

Website: www.jpathinformatics.org

DOI: 10.4103/2153-3539.179984

Quick Response Code:

This article may be cited as:
Sethi A, Sha L, Vahadane AR, Deaton RJ, Kumar N, Macias V, *et al*. Empirical comparison of color normalization methods for epithelial-stromal classification in H and E images. J Pathol Inform 2016;7:17.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2016/7/1/17/179984

INTRODUCTION

Significant advances have been made over the last decade in computational pathology – use of computer software and hardware for analysis of tissue images. Promising results have been reported not only for automated grading of cancer tissue but also for obtaining more accurate prognosis using computer algorithms.[1,2] A common step in many computational pathology pipelines is segmentation of images into different tissue compartments followed by classification using appropriate labels such as epithelium, stroma, and nuclei. Segmentation and classification algorithms frequently utilize color features, and in many cases, these algorithms are trained on hand-labeled regions delineating these compartments.

Some color variation can be useful in classifying images because such variation might reflect important contrasts in the underlying biochemical composition of the tissue. However, images of similar tissues that are colored using the same stain also suffer from unwanted color variation due to differences in stain manufacturing processes across vendors, staining protocols across labs, and color responses across digital scanners. This is especially true for the H and E stain that is universally used by surgical pathologists to reveal histopathological detail. Hematoxylin itself is a natural product extracted from logwood trees; standardization across batches is, therefore, difficult and the dye is prone to precipitation in storage, which can cause day-to-day variation even within a single lab.[3-5] In addition, the handling of the specimen during fixation and processing can alter the way in which the tissue interacts with the dyes, producing extraneous variation even in tissue microarray (TMA) cores stained on the same slide. Figure 1 shows an example of this diversity of stain appearances, among prostate cancer TMA cores scanned on the same digital microscope. The two cores shown in the left half of Figure 1 show extremes of epithelial appearance whereas the two cores in the right half show extremes of stromal appearance. Such variation in stain appearance can be problematic for algorithms in computational pathology that rely on tissue color.

While the impact of color normalization on the ultimate results of analysis will vary depending upon the specific end goal of each application, its impact on the fundamental step of epithelial-stromal segmentation is of considerable general interest and is not yet well understood. In this work, we focused on epithelial-stromal classification because it is a common early step in computational pathology, irrespective of the highly specific end goals such as cancer grading or prognosis. Separation of cancerous epithelial cells from surrounding stroma is particularly important in many analytic pipelines. We explored the advantages and

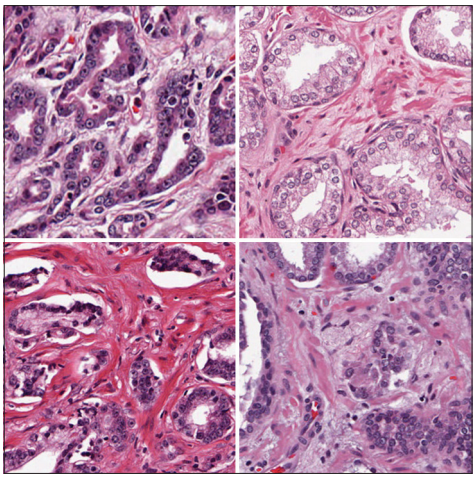

Figure 1: Diversity of H and E stained images illustrated using four prostate cancer samples with Gleason Grade 3. The first two samples show range of epithelial brightness, and the last two show the range of stromal brightness

disadvantages of two state-of-the-art color normalization techniques when used before applying a machine learning approach to classify epithelium and stroma.[4,5] The success of a machine learning approach is dependent upon the representativeness of the training data vis-à-vis the testing data, among other factors. It is natural to hypothesize that due to normalization, differences in color distribution of epithelium (or stroma) across images will be diminished, thus making it more likely that the color-normalized training set will be a good representation of the color-normalized testing set. On the other hand, there is a danger that in removing some of the inter-image color variation, some of the color differences that might be informative for further downstream objectives such as predicting prognosis might also be removed.

The two color normalization methods compared in this paper represent sophisticated contemporary methods specifically designed for analysis of H and E - stained histology images. Color normalization methods commonly used for photographic (i.e., naturally-colored) images do not take advantage of specific properties of stained tissue images. Normalization techniques meant for stained images work on each stain separately. The technique published by Vahadane *et al*. (hereafter referred to as "Vahadane") estimates sparse and nonnegative stain density maps from the color images, and then combines the stain density maps of source images with the color basis of a target image.[4] The target image usually is one with a stain appearance that is advantageous in some way. For example, it could be an image that is preferred by a pathologist for its appearance. Similarly, the technique published by Khan *et al*. (hereafter referred to as "Khan") also normalizes a source image to the color appearance of a target image.[5] It does this by estimating the color basis of both images, followed by color deconvolution and

nonlinear mapping of the source color space to match the statistics to that of the target. Color basis estimation followed by deconvolution is conceptually similar to stain separation used by Vahadane, but the two methods differ in important details. For example, stain color basis estimation in Vahadane is completely unsupervised, while the estimation of color basis in Khan is based on the classification of pixels into the two stain classes using a pretrained classifier provided by the authors.

When these normalization algorithms were first published their performance was presented only from the point of view of image appearance. These algorithms have not been tested empirically for assisting downstream goals in computational pathology such as epithelial-stromal classification. In the present work, we sought to compare the two color-normalized image sets and the original images to determine which set of source images provides the most accurate classification in an unbiased machine learning approach. We also sought to characterize the effects of each normalization technique on critical aspects of color variation that could be significant in computational pathology.

MATERIALS AND METHODS

Tissue Samples and Digitization

We used an H and E stained set of TMA slides from the cooperative prostate cancer tissue resource (CPCTR).[6] This TMA set includes quadruplicate 0.6 mm diameter core samples of prostate cancer from 404 patients and includes tumors with Gleason Grades 3, 4, and 5, and combinations thereof. The tissues were provided from five hospitals participating in the CPCTR network. Sections from the five blocks comprising the TMA set were stained with H and E at the University of Illinois at Chicago. The stained slides were scanned at ×200 on an Aperio ScanScope CS® (Leica Biosystems, Inc., Vista, CA, USA). From >1000 core images in the TMA set, we selected thirty that had significant variation in stain appearance.

Image Set Generation

Under the supervision of a pathologist, epithelial and stromal areas were separately marked on the thirty images using Aperio ImageScope® software to delineate regions serving as ground truth for both training and testing the epithelial-stromal classifiers. Due to the complex architecture of epithelial glands, it would have been impractically tedious to annotate entire images. Thus, we annotated only a few sub-regions in each image such that a large and diverse set of training examples was created. As illustrated in Figure 2, the set of thirty cores was further divided into twenty cores that were used for training epithelial-stromal classifier, and ten that were used for testing the classifiers. Six out of the ten testing images were taken from a block that was not represented in the training set at all. This simulated the real-world scenario in which software can be used on images from a lab that did not contribute data to the training of the software.

A pathologist selected a core that seemed neither over-stained nor under-stained to serve as a target for color normalization by the two methods. Thus, we

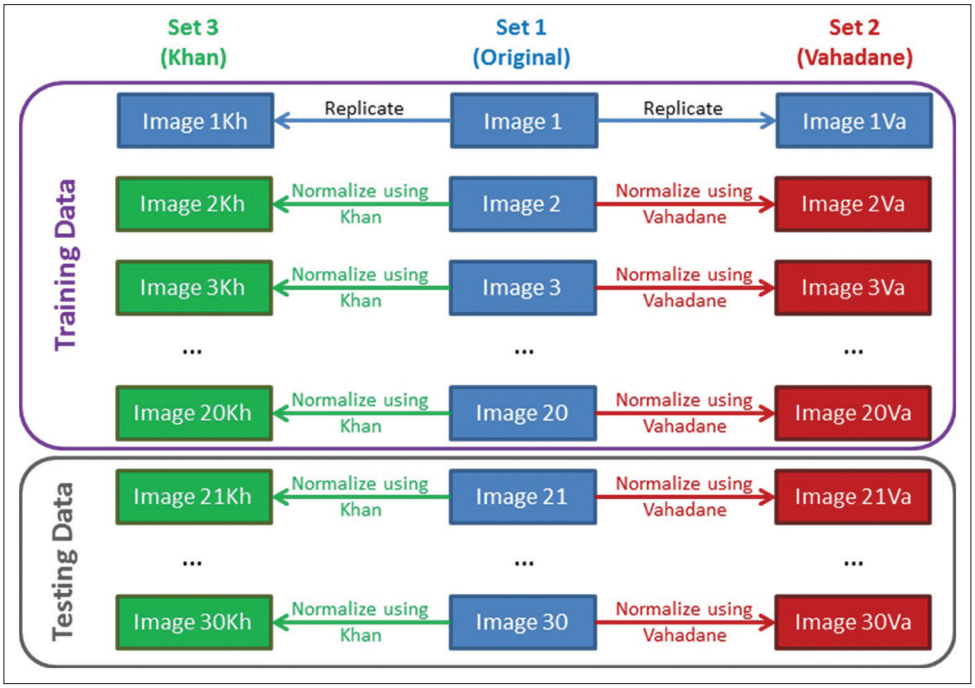

Figure 2: Preparation of the training and testing sets using original and color-normalized images

prepared two new sets of thirty images each, such that the original target image was also included in the two new sets while the 29 other images in the original set had color-normalized versions in the two new sets. Color normalization was performed using software provided by the authors of the respective techniques. Further, the ground truth regions that were marked on the thirty original images were also applied as an overlay to the corresponding images in the two normalized image sets. As shown in Figure 2, each set of thirty images included twenty images for training that included the target image, and ten images for testing that excluded the target image.

Classification Methods

We compared the efficacy of the two color normalization methods to the original images using three different procedures for epithelial-stromal classification. The main analyses were based on a method that segmented TMA core images into super-pixels – irregular clusters of pixels sharing common color and texture characteristics – and used logistic regression to classify them into epithelium or stroma. To determine whether results could be generalized to other classification methods, we also performed comparative analyses using convolutional neural networks (CNNs) and *Wndchrm*. These methods classify each pixel based on an image patch of predetermined dimensions. Since a pixel contains too little information by itself (only R, G, B values) some surrounding spatial context is needed for accurate epithelial-stromal classification. For example, both epithelial and stromal nuclei are dark in H and E stained images, but stromal nuclei are surrounded by pink-colored stroma while the epithelial nuclei are not. To classify a given pixel in the context of its spatial neighborhood, one can use either a super-pixel containing that pixel or a patch centered at that pixel. The difference between these two approaches is as follows. While the super-pixels of an image are nonoverlapping, contiguous, and irregularly-shaped subsets of its pixels, the set of patches centered at each pixel comprises fixed-sized squares that overlap with other patches centered at neighboring pixels.

In the first classification method, a technique called multiresolution segmentation (MRS) was used to obtain super-pixels.[7] This was followed by the extraction of a fixed set of features for each super-pixel. An L1-regularized logistic regression classifier was trained on labels (epithelium or stroma) of annotated super-pixels with known classes from training images. All the pixels of an annotated super-pixel were deemed to belong to the same class as the super-pixel. The results were analyzed at both super-pixel and pixel levels.

The second technique was *Wndchrm*, which is easy-to-use trainable image classification software created by Shamir *et al*. under funding from the National Institutes of Health/National Institute of Aging for biologists with no programming background.[8] It automatically extracts between 1,000 and 3,000 predefined image features from an image and trains a variation of nearest neighbor classifier on image class labels. In a limited but diverse set of image recognition tasks, it has shown performance close to or better than other state-of-the-art methods. It has an open-source software implementation. We adapted its use for classification of annotated pixels in training and testing images by using the patches centered at these pixels as input, and their annotated class as desired output.

The third technique was a deep learning technique based on CNNs, which take an image patch as input. CNNs are multi-layered neural networks in which certain layers have connectivity and weight-sharing constraints that can be implemented as a convolution operation that mimics location-invariant feature extraction of the mammalian visual cortex.[9] In addition to producing state-of-the-art results for recognition of small images, CNNs have also been used with much success in mapping regions of certain classes in large images using a patch-based approach such as detection of mitotic nuclei in whole slides.[10] The defining feature of CNNs is co-learning of a data-driven and task-specific hierarchical set of features along with a classifier using the spatial structure of the input images or patches instead of learning only a classifier on a fixed and hand-picked set of features.

Data Preparation, Training, and Testing for the Super-pixel-based Classifier

Using Definiens Developer XD® (Definiens, Munich, Germany), we segmented each image into super-pixels without any constraints on their shape. We excluded all pixels with brightness above a certain threshold as whitespace. A key feature of this software is an ability to create hierarchical layers of object maps such that sub-objects in a lower layer are mutually exclusive and collectively exhaustive subsets of the objects in the upper layer. In histopathology, this can capture relations such as various tissue compartments in an upper layer, and their nuclei and cytoplasm in a lower layer. We segmented the remaining super-pixels further into sub-objects based on their brightness relative to their neighboring pixels and area, calling the darker ones that were more than 15 pixels in area DarkSmall. DarkSmall sub-objects usually corresponded to nuclei in H and E stained images.

We labeled super-pixels that had more than 95% of the constituent pixels belonging to the hand-annotated epithelial regions as epithelium. We labeled stroma super-pixels similarly, based on >95% overlap with annotated stromal regions. In other words, if we denote the set of pixels annotated as epithelium as *AnnotatedEpi*, the set of pixels annotated stroma as *AnnotatedStroma*, the set of pixels in i^{th} super-pixel as $SuperPixel_i$, then the

decision criteria for labeling each super-pixel for training and testing can be described as follows:

$$\text{Label}\,(SuperPixel_i)=\begin{cases}\text{Epi}, & \text{if}\ \dfrac{|SuperPixel_i \cap AnnotatedEpi|}{|SuperPixel_i|}>0.95\\ \text{Stroma}, & \text{if}\ \dfrac{|SuperPixel_i \cap AnnotatedStroma|}{|SuperPixel_i|}>0.95\\ \text{Unused}, & \text{otherwise}\end{cases}$$

Thus, the selection of training super-pixels was automated using the same selection rule for each image type. This avoided a bias that could be introduced by human selection of the training images or super-pixels. We then exported a total of 93 features for every super-pixel labeled epithelium or stroma in the training images. Six types of features were exported: (1) Color, (2) texture, (3) shape, and (4) size metrics of super-pixels, (5) their relative appearance compared to neighboring super-pixels, and (6) relative features of the dark sub-objects within a super-pixel. A complete list of exported features is given in the supplementary material Table S1. Each of the training subsets contained between 5,000 and 6,000 super-pixels of which around 40% were epithelium, and the rest were stroma. Values of each feature were normalized to have zero mean and unit variance.

Using R Development Core Team,[11] we trained separate logistic regression models on the three image sets. The models were regularized using an optimization penalty on the L1-norm of their coefficients. If the weight λ of the L1-norm penalty is increased, it not only reduces the magnitude of the coefficients of all features, but it can also drive some coefficients to be exactly zero.[12] However, the latter was not our intent as we empirically observed that trying to get a sparse model in this manner reduced the model accuracy as it also shrank coefficients of useful features. Hence, we only used a light penalty ($\log \lambda = -1$) for regularizing the model to boost its validation performance while reducing features in a step-wise manner similar to backward elimination. Usually, an L2-norm penalty is used for model regularization when variable elimination is not desired, but we observed slightly higher validation accuracies when L1-norm penalty was used instead. For each of the three training sets, we started with all variables and eliminated that variable with each step which least reduced the area under receiver operating characteristic (ROC) curve area under the curve (AUC) on super-pixels using cross-validation.

We examined the classification of both super-pixels as well as individual pixels for the ten test images from each of the three image sets. For testing super-pixel classification, we analyzed the performance of a model learned using super-pixels labeled epithelium or stroma from the twenty training images in a particular image set on labeled super-pixels in the ten test images from the same set. Models with widely varying numbers of features were compared.

For testing pixel-level classification, we plotted the test-AUC of super-pixel classification versus number of variables for all three image sets and selected the number of variables that represented an "elbow", that is the minimum number of variables that could avoid a sharp decrease in discrimination capacity. These three models were built into separate Definiens Developer® rulesets and applied to the ten test images for their corresponding image sets based on the same segmentation and feature calculation process as the one used for generating the training data. The value of the logistic regression formula for a nonwhite space super-pixel was compared against a threshold for assigning an epithelium or stroma label. All constituent pixels of a super-pixel were assigned its label.

We computed sensitivity and specificity for different thresholds to plot ROC curves for the three respective test image sets. The sensitivity referred to the proportion of hand-annotated epithelial pixels across all test images that were correctly identified by the formula and the threshold. Similarly, specificity referred to the proportion of correctly identified hand-annotated stromal pixels.

In other words, the true positive rate (TPR) and true negative rate (TNR) can be expressed in terms of the set of pixels that belong to i^{th} epithelial super-pixel $EpiSuperPixel_i$, j^{th} stromal super-pixel $StromaSuperPixel_j$, the set of annotated epithelial pixels *AnnotatedEpi*, and the set of annotated stromal pixels *AnnotatedStroma* as follows:

$$\text{TRP}=\frac{\left|\cup_i (EpiSuperPixel_i \cap AnnotatedEpi)\right|}{|AnnotatedEpi|}$$

$$\text{TNR}=\frac{\left|\cup_j (StromaSuperPixel_j \cap AnnotatedStroma)\right|}{|AnnotatedStroma|}$$

Data Preparation, Training, and Testing for the Patch-based Classifiers

Due to the relative abundance of stromal pixels compared to epithelial pixels in the prostate cancer images, there is a chance to bias classifiers. Therefore, we trained and tested on an equal number of samples from the two classes by undersampling stromal pixels. Training and testing data set preparation for CNN- and *Wndchrm*-based classification was straightforward. To prepare the training data, we uniformly sampled patches of size 31 × 31 each from annotated epithelium and stroma regions from the twenty training images and

assigned the annotation as the target class. Similarly, we prepared a test set with equal number of samples from both classes.

Using patches centered at sampled annotated pixels as input, we trained a CNN to make a binary decision between epithelium and stroma. We coded CNN architecture, its training, and its testing algorithms in Theano based on python.[13] The architecture that we used had two convolution and pooling layers, and two fully connected layers. The first and second convolution layers had twenty and forty kernels (filters), respectively, with rectified linear unit nonlinearity. Max pooling of size 2 × 2 was used in both pooling layers. The two fully connected layers had 400 neurons with tanh nonlinearity. It was followed by the two softmax output nodes, one for epithelial, and other for stroma. A dropout of 0.1, 0.2, and 0.5 was used in the convolution 1, convolution 2, and fully connected layers, respectively. This architecture is similar to the one used to detect mitotic nuclei, and CNN performance is known to be robust to minor changes in architecture.[10] We tested the trained CNN on 100,000 annotated pixels and their surrounding patches of size 31 × 31 extracted from ten test images.

To train and test *Wndchrm*, we used the default settings given in the software. *Wndchrm* does not require as much training data as CNN. Therefore, we used 10,000 patches for training and 2,000 for testing with equal representation from each class and each image.

RESULTS

Impact of Color Normalization on Color Metrics

The R, G, and B color channels are highly correlated in H and E stained images. Therefore, it is more insightful to examine the distribution of pixels in the HSI space, where H refers to hue, S refers to saturation, and I refer to intensity (brightness). This is closer to how humans perceive color. We expected color normalization to reduce inter-image variance of color measures. Across 30 images, as seen in Figure 3, variances of mean pixel intensity and saturation for each image were reduced significantly in images color-normalized using Khan. Moreover, the saturation was much lower for Khan, especially in epithelium. On the other hand, Vahadane reduced the variance of hue significantly while leaving the variance of intensity and saturation almost same as the original. This implies that while the brightness and saturation were normalized by Khan method, Vahadane mainly normalized hue. Further, mean saturation was significantly reduced by Khan method, especially for epithelium, giving an overall grayish appearance to each image.

We observed that intra-image variance of hue was also significantly reduced using Vahadane method, as shown

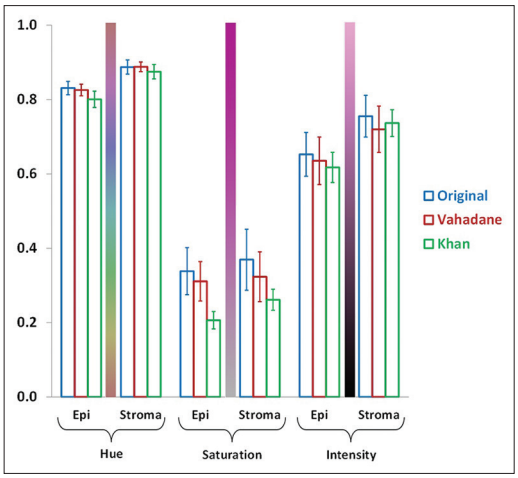

Figure 3: Color normalization illustrated using inter-image standard deviation (error bars) of mean (bars) hue, saturation, and intensity for epithelium and stroma. Continuous color bars between epithelium and stroma illustrate the hue, saturation, and intensity range holding the other two at their means

in Figure 4. This is in line with Vahadane's interpretation of color-normalization, which leaves intensity variations intact within each image by preserving their stain density maps while standardizing their RGB proportions, which determines hue. On the other hand, improvement in epithelial-stromal classification after applying Khan's color normalization can be attributed to the increase in the difference between mean epithelial and stromal intensities. These effects can also be seen in sample images in Figure 5.

Figure 6 shows an example of a part of an image on the left. In the middle, its segmentation into super-pixels using MRS on R, G, and B channels with no constraint on shapes, and a scale parameter of 75 is shown. The panel on the right shows its segmentation of DarkSmall sub-objects.

Impact of Color Normalization on Epithelial-stromal Classification

Test set AUCs for super-pixel classification using logistic regression classifiers with varying number of features for the three image sets are shown in Figure 7. Vahadane and Khan showed similar performance with no appreciable decrease until around twenty features. On the other hand, the classifiers trained on original variables showed decreased performance when the features were reduced below eighty.

Although the performance of models trained on the three sets of images was similar beyond eighty variables, computation of class labels for super-pixels becomes slower and thus less efficient with a large number of variables. Therefore, for pixel-level classification, models with twenty variables, each was selected for the three image sets. Classifier performance on the two sets of normalized images was nearly constant for twenty or more

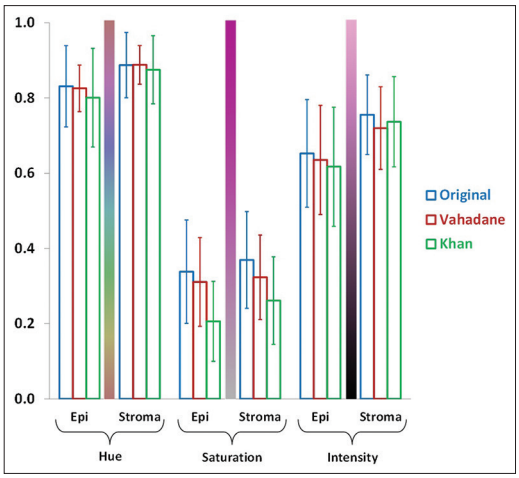

Figure 4: Contribution of color to epithelial-stromal classification illustrated using mean intra-image standard deviation (error bars) around mean (bars) hue, saturation, and intensity. Color bars between epithelium and stroma illustrate the full range of hue, saturation, and intensity while holding the other two constant

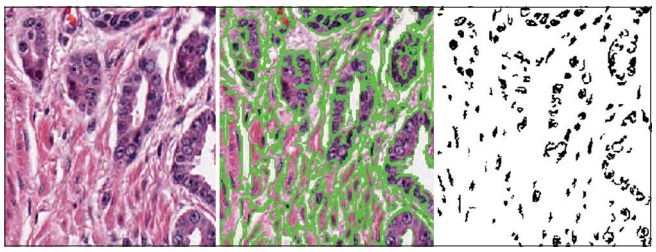

Figure 6: An example H and E stained tissue image (left), its super-pixel boundaries (center, green), and detected dark sub-objects (right, black)

variables while no such "elbow" was evident for original images. Pixel classification sensitivity and specificity on the ten test images for different thresholds are shown in Figure 8 for the three sets, which form partial ROC curves. Relative performance at pixel-level was similar to super-pixel-level performance in that Vahadane and Khan were very close and performed marginally better than the original set of images.

Twenty variables that survived backward elimination for the three image sets are listed in Table 1. The variables are first sorted by prevalence such that those that appear in all three sets appear toward the top, and then by sum rank across the three sets. Average red value (mean layer 1) was the most important feature among all three models, as expected due to the characteristic redness of stroma. Interestingly, red color difference to neighboring super-pixels also played an important role in all three models indicating more variance of red channel (mainly brightness) in epithelium than in stroma. Both these variables had the same rank in all three models. There were nine other variables that appeared in more than one model with similar ranks, indicating a reassuring degree of model concordance. The model for original images did

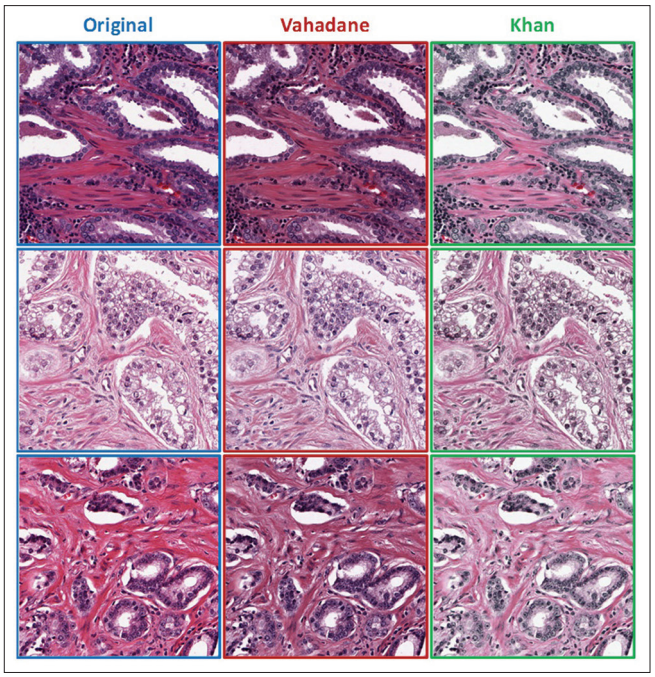

Figure 5: Selected sub-images and their normalized versions

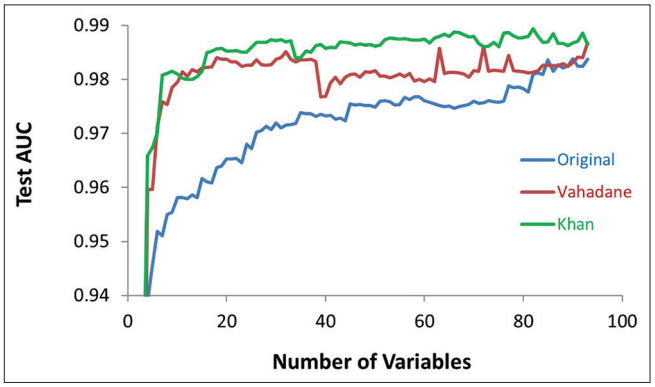

Figure 7: Test - area under the curve for the three sets of images for logistic regression models using different number of variables

not depend on features from dark sub-objects (nuclei) while the model for Khan images did not depend on texture features based on gray-level co-occurrence matrix.

To confirm the advantage of color normalization, we conducted another experiment. We used the twenty features that survived backward elimination for original images and tested their utility in classifying super-pixels extracted from the two sets of normalized images. As expected, the accuracy decreased slightly compared to what it would have been had we used the variables listed in Table 1 for the respective columns of normalized sets (test-AUC was 0.973 instead of 0.984, and 0.976 instead 0.985 for Vahadane and Khan methods, respectively). However, the trend still held that color-normalized training sets yielded slightly more accurate epithelial-stromal classifiers (comparable test-AUC using original images was 0.965). This suggests

that even when the variables are not specifically selected for the color-normalized datasets, they yield slightly better classification than original images.

Pixel-level classification AUC for CNN and *Wndchrm* are shown in Table 2, and these results are also generally consistent with the MRS results. Both color normalization methods improve classification accuracy, and Vahadane method seems to give relatively higher results compared to Khan method. The accuracy boost from color normalization is greater for both patch-based methods than we observed for the super-pixel method.

Some examples of resultant epithelial-stromal maps obtained using MRS and a logistic regression threshold of 0.5 on the twenty-variable models for the three sets of images are shown in Figure 9. We observed that more of the light-colored epithelium was confused as stroma in the original images [see lower example in Figure 9]. On the other hand, due to color normalization, both Vahadane and Khan had less trouble with such images, thus explaining their improved performance over the original set. Relative to Vahadane, Khan struggled with correctly identifying epithelium with nuclei whose chromatin had marginated. This is likely because Vahadane estimates the stain density maps in an unsupervised fashion. Therefore, it adapts to the images including those with light colored epithelium whereas Khan uses a pretrained stain classification model. On the other hand, Vahadane struggled relatively more with identifying inflamed stroma as stroma because of the concentration of hematoxylin in clustered lymphocytes. In this case, Khan's pretrained model seems to help in identifying the matrix surrounding lymphocytes correctly.

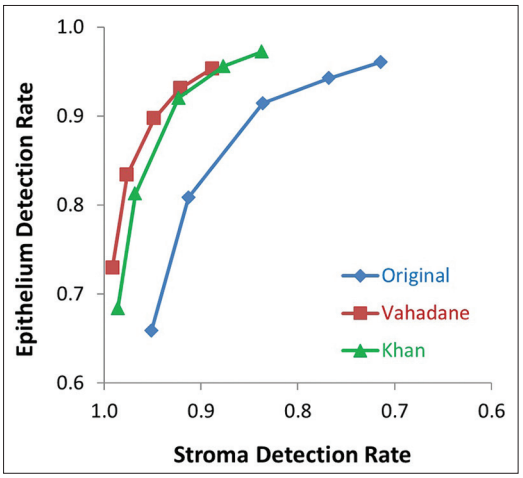

Figure 8: Receiver operating characteristic curves for pixel-level accuracy for twenty-feature models for ten test images for the three models for thresholds 0.75, 0.5, 0.25, 0.15, and 0.1 on the logistic regression output

Testing Run-time for Color Normalization Methods

Color normalization of images does take computation time, which varies by technique. In our experiments on a computer running Windows 7 with Intel Core i7 CPU that has four cores and 16GB RAM and runs at a clock speed of 3.4GHz, Vahadane method took 1136s and Khan method took 633s to color-normalized thirty images each of size 2000 × 2000 pixels. The larger time taken by Vahadane is likely due computation of a color model afresh for each image while Khan relies on a precomputed model.

DISCUSSION

This empirical analysis confirmed using multiple classification approaches that color normalization using both Vahadane and Khan methods helped in increasing the accuracy of epithelial-stromal classification; however,

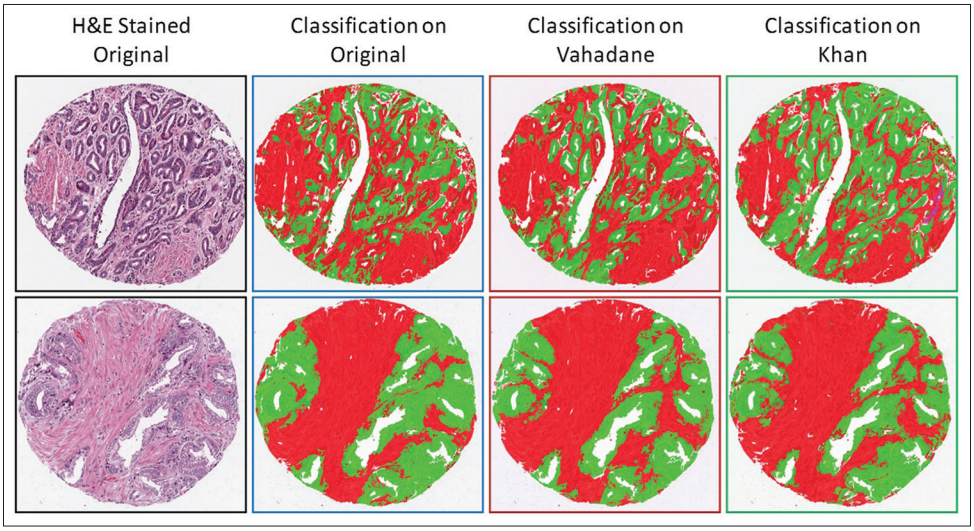

Figure 9: Two examples of cores whose pixels have been classified into epithelium (green) and stroma (red) based on original images as well as normalized images using Vahadane and Khan methods

Table 1: Features ranks based on coefficient magnitude in the 20-variable models for the three image sets (sorted by prevalence and sum rank across the three sets)

Feature	Original	Vahadane	Khan
Mean layer 1	1	1	1
Mean difference to neighbors layer 1 (0)	3	3	3
Mean difference to neighbors layer 3 (0)	4	4	
Brightness	2		7
Relative area of sub-objects DarkSmall (2)		11	8
Roundness	9		11
SD layer 2	8		12
Shape index	14		10
Compactness (polygon)		12	18
Radius of smallest enclosing ellipse	16	14	
Maximum branch length (Pxl)	18		17
GLCM mean (quick 8/11) (all direction)		2	
Mean layer 3			2
SD to neighbor pixels stain 1 (3)			4
Border length (Pxl)	5		
SD to neighbor pixels layer 1 (3)			5
SD to neighbor pixels layer 2 (3)		5	
Mean difference to neighbors stain 2 (0)			6
Number of segments	6		
SD to neighbor pixels layer 3 (3)		6	
GLCM SD (quick 8/11) (all direction)	7		
Perimeter (polygon) (Pxl)		7	
Number of edges (polygon)		8	
Contrast to neighbor pixels layer 2 (3)			9
Contrast to neighbor pixels layer 3 (3)		9	
GLCM entropy (quick 8/11) (all direction)		10	
SD layer 1	10		
SD stain 1	11		
Density	12		
Asymmetry			13
Border index	13		
SD Layer 3		13	
Skewness stain 1			14
Border contrast layer 2	15		
Skewness layer 1		15	
Width (only main line) (Pxl)			15
Degree of skeleton branching			16
GLCM Angular 2nd moment (quick 8/11) (all direction)		16	
Edge contrast of neighbor pixels (prototype) stain 2 (3)	17		
SD of area represented by segments (Pxl)		17	
Asymmetry of sub-objects SD (2)		18	
Average area represented by segments (Pxl)			19
Border contrast layer 1	19		
Skewness stain 2		19	
Area (including inner polygons) (Pxl)		20	
Border contrast stain 2			20
Radius of largest enclosed ellipse	20		

GLCM: Gray-level co-occurrence matrix, SD: Standard deviation

the extent of improvement was smaller than expected for the super-pixel-based approach. This was perhaps because the classification of original images itself was fairly accurate to begin with. We conjecture that this is due to the rich set of features extracted from the images, particularly features that encode texture and relative appearance between a super-pixel and its neighbors. Staining differences between images are likely to affect

Table 2: Classification test area under receiver operating characteristic curve for epithelial and stromal pixels based on their surrounding patches

Technique	Original	Vahadane	Khan
WNDCHRM	0.922	0.977	0.962
Convolutional neural networks	0.921	0.965	0.948

super-pixels and their neighbors in the same images in a similar manner. Therefore, texture and relative features that encode spatial variance in color ensure implicit color normalization to some extent before explicit color normalization is even applied. In addition, starting with a large set of features, eliminating the least useful features using backward elimination, and regularizing the models is likely to yield high performing classification models.

We believe that the two techniques compared in this paper represent the state-of-the-art in histological stain normalization.[4,5] Color normalization using either a standardizing stain or a color chart as a calibration slide requires an extra scan, which does not even guarantee intra-batch invariance.[14] Simpler histogram-based adjustments do not take differences in stain proportion and concentration across slides into account.[15] A more sophisticated histogram matching-based approach that first separated a given tissue into classes such as nuclei, lumen, and cytoplasm, and then matched their respective histograms across images has also been proposed, but it is not clear if such intra-class histogram matching leads to loss of differential information between the same class of objects across disease states.[16] One of the earliest stain separation techniques was proposed by Ruifrok and Johnston in which the stain color basis was extracted using control slides with single stains.[17] Khan can be seen as its extension and improvement that uses a pretrained stain model based on several samples.[5] On the other hand, Vahadane *et al*.[4] can be seen as an extension and improvement of Macenko *et al*.[18] that used singular value decomposition for separation of stain density maps, or Díaz and Romero[19] that used nonnegative matrix factorization (NMF) while Vahadane uses sparse NMF for computing the color basis from the given image itself. Recently, a similar technique independently developed by Xu *et al*. has also been published.[20]

One of the relative advantages of Vahadane over Khan was derivation of data-driven stain models instead of relying on a pretrained one. Consequently, it requires more computational time than Khan method. Such an approach can easily adapt to other stains without the need to estimate the stain models. However, given a diverse training set, a pretrained model can cover most of the variation in stain appearance that one is likely to encounter. This, in addition to the feature set and modeling procedure advantages cited above, might explain why there was very little difference between the performances of the two color normalization methods in spite of the differences in their approach. The difference between the normalization methods was evident in the appearance of normalized images as well as their color statistics, with the Vahadane method providing colors more typically encountered by pathologists.

While our simple nucleus segmentation method had false positives and fractured or clumped nuclei, it served to generate features for epithelial-stromal classification. A more sophisticated nuclear resegmentation can be used downstream that relies on the contextual knowledge thus obtained for better segmentation. For example, different shape priors for stromal nuclei (more spindle-shaped) can be implemented compared to epithelial nuclei (more round), after epithelial-stromal classification based on crude segmentation of small dark objects. This approach was used in Beck *et al*.'s image processing pipeline for breast cancer images.[2]

We varied the target image for color normalization to a more reddish one and a more bluish one, and the results were similar in that color normalization using Vahadane or Khan led to slightly better epithelial-stromal classification than the original images. Varying the test fold (set of held out images) also yielded similar results with the exception of one image with an unusually large amount of inflammation. If this image was part of the test fold, all three classifiers confused a large proportion of inflamed stroma as epithelium, with the classifier trained on Vahadane set being more notably affected than others.

To our knowledge, this is the first analysis that compares different color normalization methods for meeting a goal in computational pathology instead of their improvement in subjective appearances, although the impact of a single technique on nuclear segmentation, which is another common intermediate goal, has been studied.[21] Our framework for comparison incorporated several measures for an unbiased comparison of original and color-normalized images. For example, since application of MRS on color-normalized images leads to different super-pixel boundaries than those on the original images, we automated selection of training super-pixels based on a uniform criterion for their overlap with the same underlying hand-annotated regions. We also utilized the same hand-annotated regions to assess pixel-level accuracy. We compared the models for the three sets of images at pixel-level using the same number of variables. In addition to giving a level playing field to the three sets of images, we also tried to simulate a real-world scenario wherein software is useful only if it is applied to data from hospitals that did not contribute data for building (training) the software. To do so, we held out all super-pixels from a subset of images

for testing, including a whole block that came from a particular hospital, instead of letting super-pixels from same images to be a part of training and testing sets.

Although primary analyses were conducted using the MRS super-pixel approach, our data indicate that the same trend across original and normalized images could be observed when additional, patch-based, classification methods are employed. Color normalization induced a somewhat smaller increase in accuracy for the MRS method, presumably due to the inherent control of some color variation already built into the super-pixel approach, as described above. We attribute this to use of relative features of a super-pixel with respect to its neighbors in MRS, which can be interpreted as implicit color normalization. Absolute levels of accuracy across these classification methods – CNN and *Wndchrm* – should be interpreted cautiously since our main focus was on MRS and determining the relative effects of color normalization. It is possible that further optimization of parameters, especially for CNN, could further increase the absolute level of accuracy. For example, increasing patch size, altering architecture, and postprocessing to remove spatially isolated misclassification could conceivably further reduce classification error.

Epithelial-stromal classification for testing images taken by a different scanner than the one used for training images was not tested in our study, although both Khan and Vahadane demonstrated the ability to subjectively color-normalize H and E stained images across scanners.[4,5] There was also a large variation in epithelial-stromal classification performance across images. As mentioned previously, some of these variations were dependent on the amount of inflammation in different images.

Recently, Zarella *et al*. proposed a technique to classify pixels of H and E stained slides into different histologic structures (such as nuclei, cytoplasm, and stroma) based on agglomerative clustering and user-defined class assignment to the clusters.[22] It is hard to compare our results with theirs because they used a different set of images that are not publicly available, and reported different metrics, that is, concordance rates for different certainty levels. It seems that their definition of concordance level is the same as our definition of pixel-level accuracy, and the two are comparable in magnitude (their concordance was 0.92–0.95 while equal-error rate pixel-level accuracy for Vahadane and Khan reported in Figure 8 is 0.92). However, certainty level was not defined in their paper. Importantly, their goal was a more direct classification of pixels rather than the use of color normalization across images while ours was to compare the effect of state-of-the-art color normalization techniques.

While our results should have some general applicability for the common goal of epithelial-stromal segmentation, it is not clear how color normalization will impact downstream goals such as clinical end-point prediction. To the extent that useful information about these end-points is captured by color, for instance, variation of basophilia of different tissue compartments with the pathological condition, it is possible that color normalization would negatively affect the accurate assessment of the underlying pathology or prognosis. However, a straightforward way around this is to use color normalization for epithelial-stromal classification, but then use the classification maps thus obtained on the original images themselves. This can render moot the question as to whether the color alterations by one technique or another are acceptable, as we will only be concerned with the contribution of that technique to epithelial-stromal classification. On the other hand, it is also possible that the color variations muted by color normalization may improve estimation of the pathology itself. To address these possibilities, thorough computational pathology investigations that are specific to the desired end goals after epithelial-stromal classification need to be conducted.

CONCLUSIONS

The three epithelial stromal classifiers that were trained and tested performed slightly better on images normalized by either technique as compared to training and testing on original images. This advantage ranged from slight in case of the super-pixel-based classification method, to substantial in case of CNN and *wndchrm*. We conjecture that some of the features used in super-pixel-based classification method perform implicit color normalization, thus diminishing the advantage of preprocessing based on explicit color normalization.

Acknowledgment

The authors would like to thank Peter Nguyen and Milita Petrauskaite for their help in annotating epithelium and stroma regions under the supervision of Dr. Virgilia Macias.

Financial Support and Sponsorship

Indo-US Science and Technology Forum, R01CA155301 from the National Cancer Institute/National Institutes of Health.

Conflicts of Interest

There are no conflicts of interest.

REFERENCES

1. Tabesh A, Teverovskiy M, Pang HY, Kumar VP, Verbel D, Kotsianti A, *et al*. Multifeature prostate cancer diagnosis and Gleason grading of histological images. IEEE Trans Med Imaging 2007;26:1366-78.
2. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, *et al*.

Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci Transl Med 2011;3:108.
3. Anagnostou VK, Welsh AW, Giltnane JM, Siddiqui S, Liceaga C, Gustavson M, *et al.* Analytic variability in immunohistochemistry biomarker studies. Cancer Epidemiol Biomarkers Prev 2010;19:982-91.
4. Vahadane A, Peng T, Albarqouni S, Baust M, Steiger K, Schlitter AM, *et al.* Structure-Preserved Color Normalization for Histological Images. International Symposium on Biomedical Imaging (ISBI): From Nano to Macro, New York, USA; April, 2015.
5. Khan AM, Rajpoot N, Treanor D, Magee D. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Trans Biomed Eng 2014;61:1729-38.
6. Patel AA, Kajdacsy-Balla A, Berman JJ, Bosland M, Datta MW, Dhir R, *et al.* The development of common data elements for a multi-institute prostate cancer tissue bank: The Cooperative Prostate Cancer Tissue Resource (CPCTR) experience. BMC Cancer 2005;5:108.
7. Baatz M, Schäpe A. Multiresolution segmentation-an optimization approach for high quality multi-scale image segmentation. In: Strobl J, Blaschke T, Griesebner G, editors. Angew. Geogr. Info. verarbeitung, Wichmann-Verlag, Heidelberg. 2000. p. 12-23.
8. Shamir L, Orlov N, Eckley DM, Macura T, Johnston J, Goldberg IG. *Wndchrm* – An open source utility for biological image analysis. Source Code Biol Med 2008;3:13.
9. Lawrence S, Giles CL, Tsoi AC, Back AD. Face recognition: A convolutional neural-network approach. IEEE Trans Neural Netw 1997;8:98-113.
10. Ciresan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. Med Image Comput Comput Assist Interv 2013;16(Pt 2):411-8.
11. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2008. Available from: http://www.R-project.org. [Last accessed on 2016 March 15].
12. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Methodol 1996;58:267-88.
13. Bergstra J, Bastien F, Breuleux O, Lamblin P, Pascanu R, Delalleau O, *et al*. Theano: Deep Learning on GPUs with Python. Neural Information Processing Systems; 2011.
14. Bautista PA, Hashimoto N, Yagi Y. Color standardization in whole slide imaging using a color calibration slide. J Pathol Inform 2014;5:4.
15. Reinhard E, Ashikhmin M, Gooch B, Shirley P. Color transfer between images. IEEE Comput Graph Appl 2001;21:34-41.
16. Basavanhally A, Madabhushi A. EM-based segmentation-driven color standardization of digitized histopathology. SPIE Med Imaging 2013;8676.
17. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol 2001;23:291-9.
18. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, *et al.* A method for normalizing histology slides for quantitative analysis. Proceeding IEEE International Symposium Biomedical Imaging 2009;9:1107-10.
19. Díaz G, Romero E. Micro-structural tissue analysis for automatic histopathological image annotation. Microsc Res Tech 2012;75:343-58.
20. Xu J, Xiang L, Wang G, Ganesan S, Feldman M, Shih NN, *et al.* Sparse Non-negative Matrix Factorization (SNMF) based color unmixing for breast histopathological image analysis. Comput Med Imaging Graph 2015;46(Pt 1):20-9.
21. Monaco J, Hipp J, Lucas D, Smith S, Balis U, Madabhushi A. Image segmentation with implicit color standardization using spatially constrained expectation maximization: Detection of nuclei. Med Image Comput Comput Assist Interv 2012;15(Pt 1):365-72.
22. Zarella MD, Breen DE, Plagov A, Garcia FU. An optimized color transformation for the analysis of digital images of hematoxylin and eosin stained slides. J Pathol Inform 2015;6:33.

Table S1: Features exported for each super-pixel for training epistromal classifiers

Feature type	List of features
Color	Brightness, mean layer 1, mean layer 2, mean layer 3, mean stain 1, mean stain 2
Texture	SD layer 1, SD layer 2, SD layer 3, SD stain 1, SD stain 2, skewness layer 1, skewness layer 2, skewness Layer 3, skewness stain 1, skewness stain 2, GLCM homogeneity (quick 8/11) (all direction), GLCM contrast (quick 8/11) (all direction), GLCM dissimilarity (quick 8/11) (all direction), GLCM entropy (quick 8/11) (all direction), GLCM Angular 2nd moment (quick 8/11) (all direction), GLCM mean (quick 8/11) (all direction), GLCM SD (quick 8/11) (all direction)
Shape	Length/width, asymmetry, border index, compactness, density, elliptic fit, rectangular fit, roundness, shape index, compactness (polygon), number of edges (polygon), number of inner objects (polygon), SD of length of edges (polygon) (Pxl), curvature/length (only main line), degree of skeleton branching, length/width (only main line), number of segments, SD of area represented by segments (Pxl)
Size	Border length (Pxl), length (Pxl), number of pixels, width (Pxl), radius of largest enclosed ellipse, radius of smallest enclosing ellipse, Area (excluding inner polygons) (Pxl), area (including inner polygons) (Pxl), average length of edges (polygon) (Pxl), length of longest edge (polygon) (Pxl), perimeter (polygon) (Pxl), average area represented by segments (Pxl), length of main line (no cycles) (Pxl), length of main line (regarding cycles) (Pxl), maximum branch length (Pxl), width (only main line) (Pxl)
Relation to neighbors	Border contrast layer 1, border contrast layer 2, border contrast Layer 3, border contrast stain 1, border contrast stain 2, contrast to neighbor pixels layer 1 (3), contrast to neighbor pixels layer 2 (3), contrast to neighbor pixels layer 3 (3), contrast to neighbor pixels stain 1 (3), contrast to neighbor pixels stain 2 (3), edge contrast of neighbor pixels (prototype) layer 1 (3), edge contrast of neighbor pixels (prototype) layer 2 (3), edge contrast of neighbor pixels (prototype) layer 3 (3), edge contrast of neighbor pixels (prototype) stain 1 (3), edge contrast of neighbor pixels (prototype) stain 2 (3), SD to neighbor pixels layer 1 (3), SD to neighbor pixels layer 2 (3), SD to neighbor pixels layer 3 (3), SD to neighbor pixels stain 1 (3), SD to neighbor pixels stain 2 (3), mean difference to neighbors layer 1 (0), mean difference to neighbors layer 2 (0), mean difference to neighbors layer 3 (0), mean difference to neighbors stain 1 (0), mean difference to neighbors stain 2 (0)
Relation to sub-objects	Area of sub-objects mean (2) (Pxl), area of sub-objects SD (2) (Pxl), density of sub-objects mean (2), density of sub-objects SD (2), asymmetry of sub-objects mean (2), asymmetry of sub-objects SD (2), direction of sub-objects mean (2), direction of sub-objects SD (2), area of sub-objects DarkSmall (2) (Pxl), Number of sub-objects DarkSmall (2), Rel. area of sub-objects DarkSmall (2)

SD: Standard deviation, GLCM: Gray-level co-occurrence matrix