# A Resampling-Based Markovian Model for Automated Colon Cancer Diagnosis

Erdem Ozdemir, Cenk Sokmensuer, and Cigdem Gunduz-Demir*, *Member, IEEE*

***Abstract*—In recent years, there has been a great effort in the research of implementing automated diagnostic systems for tissue images. One major challenge in this implementation is to design systems that are robust to image variations. In order to meet this challenge, it is important to learn the systems on a large number of labeled images from a different range of variation. However, acquiring labeled images is quite difficult in this domain, and hence, the labeled training data are typically very limited. Although the issue of having limited labeled data is acknowledged by many researchers, it has rarely been considered in the system design. This paper successfully addresses this issue, introducing a new resampling framework to simulate variations in tissue images. This framework generates multiple sequences from an image for its representation and models them using a Markov process. Working with colon tissue images, our experiments show that this framework increases the generalization capacity of a learner by increasing the size and variation of the training data and improves the classification performance of a given image by combining the decisions obtained on its sequences.**

***Index Terms*—Automated cancer diagnosis, cancer, histopathological image analysis, Markov models, resampling.**

## I. Introduction

COLORECTAL cancer is one of the most common yet most curable cancer types in western countries. Its survival rates increase with early diagnosis and selection of a correct treatment plan, for which correct grading is critical [1]. The final diagnosis and grading of colorectal cancer is based on histopathological assessment of biopsy tissue samples. In this assessment, pathologists decide on the presence of cancer based on the existence of abnormal formations in a tissue and determine cancer grade based on the degree of the abnormalities. As this assessment mainly relies on visual interpretation, it may contain subjectivity [2]. Thus, it has been proposed to use computational methods that help decrease the subjectivity level by providing quantitative measures.

Manuscript received April 19, 2011; revised August 5, 2011 and October 19, 2011; accepted October 23, 2011. Date of publication October 27, 2011; date of current version December 21, 2011. This work was supported by the TÜBİTAK under Project 110E232. *Asterisk indicates corresponding author.*

E. Ozdemir is with the Department of Computer Engineering, Bilkent University, Ankara TR-06800, Turkey (e-mail: erdemo@cs.bilkent.edu.tr).

C. Sokmensuer is with the Department of Pathology, Hacettepe University Medical School, Ankara TR-06100, Turkey (e-mail: csokmens@hacettepe.edu.tr).

*C. Gunduz-Demir is with the Department of Computer Engineering, Bilkent University, Ankara TR-06800, Turkey (e-mail: gunduz@cs.bilkent.edu.tr).

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

Digital Object Identifier 10.1109/TBME.2011.2173934

The previous methods provide automated classification systems that use a set of features to model the difference between the normal tissue appearance and corresponding abnormalities. These features are usually defined by the motivation of mimicking a pathologist, who uses morphological changes in cell nuclei and organizational changes in the distribution of tissue components to detect abnormalities. Morphological methods aim to model the first kind of these changes by extracting features that quantify the size and shape characteristics of cell nuclei. These features can be used to characterize an individual nucleus [3] as well as an entire tissue by aggregating the features of its nuclei [4]. Extraction of morphological features requires determining the exact locations of nuclei beforehand, which is, however, very challenging for histopathological tissue images due to their complex natures [5].

Structural methods are designed to characterize topological changes in tissue components by representing the tissue as a graph and extracting features from this graph. In literature, almost all methods construct their graphs considering nuclear components as nodes and generating edges between these nodes to encode spatial information of the nuclear components. The studies use different graph generation methods including Delaunay triangulations (and their dual Voronoi diagrams) [6], [7], minimum spanning trees [4], [8], probabilistic graphs [9], [10], and weighted graphs [11]. To model topological tissue changes better, we have recently proposed to consider different tissue components as nodes and construct a color graph on these nodes, in which edges are colored according to the tissue type of their end points [12]. Likewise, the main challenge of defining structural features is the difficulty of locating the components. The incorrect localization may affect the success of the classification systems.

Textural methods avoid difficulties relating to correct localization of cells (and other components) by defining their textures on pixels, without directly using the tissue components. They assume that abnormalities from the normal tissue appearance can be modeled by texture changes observed in tissues. There are many ways to define textures for tissues; they include using intensity/color histograms [13], co-occurrence matrices [14], [15], run-length matrices [16], multiwavelet coefficients [17], local binary patterns [18], [19], and fractal geometry [13], [20]. Textural features typically characterize small regions in tissue images well but they may have difficulties to find a constant texture characterizing the entire tissue. To alleviate this difficulty, it is proposed to divide the image into grids, compute textural features on the grids, and aggregate the features for characterizing the tissue [14]. Although grid-based approaches improve accuracies, they may still have difficulties arising from the

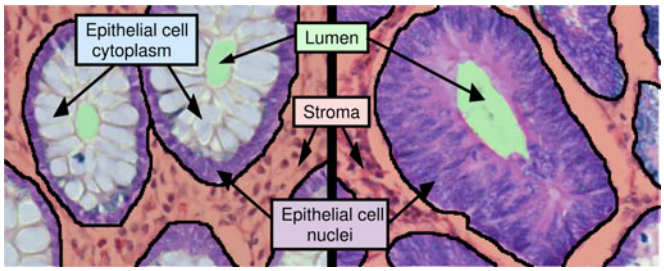


Fig. 1. Cytological components in normal and cancerous colon tissues. Different components are illustrated with different colors: green for luminal regions, red for stromal regions, purple for epithelial cell nuclei, and blue for epithelial cell cytoplasms. Colon glands are confined with black boundaries.

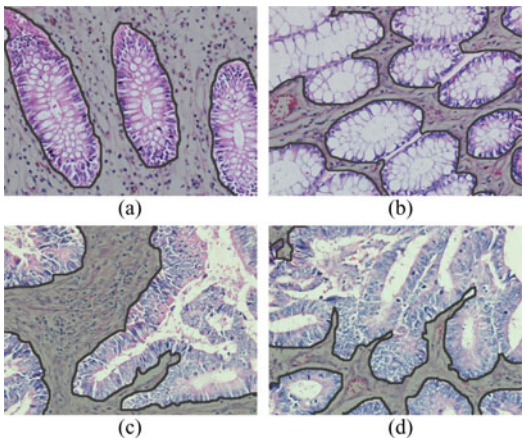


Fig. 2. Histopathological images of colon tissues: (a), (b) Normal and (c)-(d) cancerous. Nonglandular regions in images are shaded with gray.

existence of irrelevant tissue regions. For example, for diagnosing colon adenocarcinoma, which accounts for 90–95 % of colorectal cancers, pathologists examine glandular tissue regions since this cancer type originates from glandular epithelial cells and causes deformations in glands (Fig. 1). Nonglandular regions, which do not include epithelial cells, are irrelevant within the context of colon adenocarcinoma diagnosis. Moreover, such nonglandular regions can be of different sizes (Fig. 2). Thus, directly including these regions into texture computation may result in lower accuracies [21]. Aggregation methods that consider the existence of such irrelevant regions have potential to give better accuracies.

Additionally, all classification systems face a common difficulty regardless of their feature types: large variance observed in tissue images. This is mainly because of the variation among different biopsies. The variance becomes even larger due to nonideal steps in tissue preparation. Thus, to make successful generalizations [22], a classification system usually needs a large number of images from different patients in its training. On the other hand, this number is usually very limited since acquiring a large number of labeled tissue images from a large number of patients is quite difficult in this domain. When such limited data are used for training, the learned systems may be vulnerable to variations in tissue images, also leading to unstable classifications.

In this paper, we propose a new framework for the effective and robust classification of tissue images even when only limited data are available. In the proposed framework, our main contributions are the introduction of a new resampling method to simulate the variations in tissue images for learning better generalizations and the use of this method for obtaining more stable classifications. The resampling method relies on generating multiple sequences from an image, each of which corresponds to a "perturbed sample" of the image, and modeling the sequences using a first order discrete Markov process. Working with colon tissue images, our experiments show that such a resampling method is effective in increasing the generalization capacity of a learner by increasing the size and variation of the training set as well as boosting the classifier performance for an unseen image by combining the decisions of the learner on multiple sequences of that image.

This study differs from the previous tissue classification methods in two main aspects. First, it proposes a new framework in an attempt to alleviate an issue of having limited labeled training data. For that, it introduces the idea of generating perturbed images from the training data and modeling them by a Markov process. Although the issue of having limited training data is acknowledged by many researchers, it has rarely been considered in the design of tissue classification systems. Second, it proposes to classify a new image using its perturbed samples. The use of different samples of the same image is more effective to reduce the negative outcomes of large variance observed in tissue images, as opposed to the use of the entire images at once. Moreover, modeling the perturbed samples with Markov processes provides an effective method in modeling the irrelevant regions.

There also exist resampling techniques in machine learning literature. In the first category, random sampling methods, such as bootstrapping, are used especially for balancing unbalanced datasets [23]. However, such methods select new samples from the original data without changing their contents. Thus, they do not increase the variability of a training set although they can increase the size. In the second category, there exist resampling methods, such as jittering and perturbation, that help increase the variability. These methods obtain samples slightly modifying the original data [24]. The resampling method proposed by this study can be considered as an example of the latter category. It introduces a framework that modifies (perturbs) the image content to increase the data variability.

## II. Methodology

The proposed resampling-based Markovian model (RMM) relies on generating perturbed samples (sequences) from an image and using them in learning and classification. It includes two main parts: sequence generation and Markov modeling.

### A. Sequence Generation

Let $I$ be a tissue image that is to be either classified or used in training. The RMM represents this image by $N$ of its perturbed samples, $I = \{S^{(n)}\}_{n=1}^{N}$, each of which is represented by a sequence of $T$ observation symbols, $S^{(n)} = O_1^{(n)} O_2^{(n)} \ldots O_T^{(n)}$.

(For better readability, we will drop $n$ from the terms unless its use is necessary. Thus, each sample is represented by $S = O_1 O_2 \ldots O_T$.)

The first step of generating a sequence $S$ from the image $I$ is to select $T$ random data points from the image and characterize them by extracting features. The RMM proposes a generic framework that does not impose any particular feature type. Thus, one can use his/her own features within this framework to characterize the data points. In this work, we characterize each point by using pixels of its neighborhood. To this end, we locate a window at the center of each point and extract four simple features that quantify color distribution and texture of the pixels falling within this window. These four features are defined on the quantized pixels. The k-means algorithm is used to quantize the pixels of the image $I$ into three, each of which corresponds to one dominant color (white, pink, or purple) in a tissue stained with hematoxylin-and-eosin. The first three features are the ratios of these colors over the window. The last feature is a texture descriptor ($J$-value) that quantifies how uniform the quantized pixels are distributed in space [25].

After selecting the data points and extracting their features, the second step is to discretize the features into $K$ observation symbols since discrete Markov models are used. For that, we use k-means clustering to learn $K$ clusters on the features of the data points selected from the training images[1]. Then, for a new data point $P$, we use the label of the clustering vector (observation symbol $O$) whose features are the closest to those of the data point. At the end of this step, each sample is represented with a set of observation symbols, but not as a sequence of them. Thus, the next step is to order the data points and construct a sequence from their observation symbols.

The data points are so ordered as to minimize the distance between the adjacent points. Formally, this ordering problem can be represented as finding $S = O_1 O_2 \ldots O_T$ such that

$$\sum_{t=2}^{T} \text{dist}(P_{t-1}, P_t) \tag{1}$$

is minimized. Here $dist(u, v)$ is the Euclidean distance between the points $u$ and $v$ and $O_t$ is the observation symbol defined for the point $P_t$. This problem corresponds to finding the shortest Hamiltonian path among the given points, which is known as NP-complete. Thus, we use a greedy solution for ordering. This solution selects the point closest to the top-left corner as the first data point $P_1$ and then, at every iteration $t$, it selects the data point $P_t$ that minimizes $dist(P_{t-1}, P_t)$. Note that it is possible to obtain the orders using different methods. For example, one can construct a graph on the selected points based on proximity and obtain a seriated graph using Fiedler vectors [26]. Although such methods may give better sequence orders, they typically have higher computational requirements. The appendix gives the pseudocode for observation symbol learning and sequence generation.

[1]We learn clusters selecting 100 random data points from each training image. Although the number of selected points does not have too much effect for larger training sets, its smaller values lead to decreased performance when smaller training sets are used. In general, this number should be selected large enough so that different "good" clusters can be learned. However, it should be selected smaller to decrease the computational time of training.

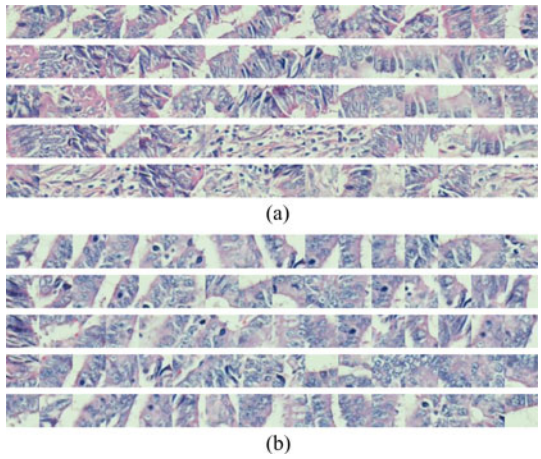


Fig. 3. Sequences generated for the tissue images given in Fig. 2(c) and (d).

At the end of this step, we have obtained $N$ sequences, which are expected to model variances in tissue images better. To illustrate the reason behind this, let us consider the images shown in Fig. 2(c) and (d) to belong to the training and test sets, respectively (we will refer to these images as `I2c` and `I2d`). In the RMM, instead of considering `I2c` as an individual training instance, we generate multiple sequences from `I2c` and put all these sequences in the training set. Fig. 3(a) illustrates five such sequences; here a data point is represented with its window, in which its features are extracted. Likewise, instead of considering `I2d` as an individual test instance, we generate multiple sequences from `I2d`, classify them by the Markov models, and combine the class of each sequence by voting. Fig. 3(b) illustrates five sequences to be classified. Now suppose that our model works on entire images but not sequences. In this case, since the training image `I2c` and the test image `I2d` show differences at the pixel level, the classifier, which was learned on the training set that includes `I2c`, may give an incorrect classification for `I2d`. Next suppose that our model works on sequences. In this case, it is more likely to correctly classify the sequences of `I2d` (and thus `I2d`) thanks to the existence of the first three sequences of `I2c` in the training set. Note that this process may generate some noisy sequences that introduce erroneous data for training. However, since the sequence length is typically selected as large, we expect the sequences to contain only partial noises. Moreover, as the RMM uses multiple sequences but not a single one, we expect this kind of erroneous sequences to be tolerated by the others provided that a large number of sequences are generated.

### *B. Markov Modeling*

The classification of a given image $I$ is done using its sequences. For each sequence $S$, the posterior probability of every class $C_m$ is computed and the class $C^*$ that maximizes these posterior probabilities is selected.

$$C^* = \underset{m}{\text{argmax}}\, P(C_m \mid S). \tag{2}$$

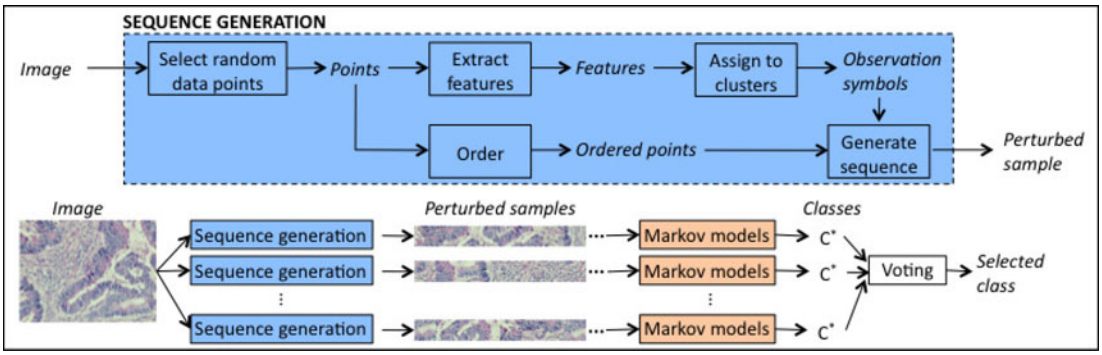


Fig. 4. A schematic overview of the proposed resampling-based Markovian model (RMM) for classifying a given image.

Subsequently, a majority voting scheme is used to combine the selected classes of the sequences.

Posteriors $P(C_m \mid S)$ are estimated using first order discrete Markov models; it is assumed that there exist dependencies between subsequent observation symbols and that there is one-to-one correspondence between observation symbols and states. Thus, in the proposed RMM, the states are observable and each sequence $S = O_1 O_2 \ldots O_T$ satisfies the Markovian property, in which the current state (observation symbol) depends on only its predecessor state.

$$P(O_t = v_i \mid O_{t-1} = v_j, O_{t-2} = v_k, \ldots) = P(O_t = v_i \mid O_{t-1} = v_j). \quad (3)$$

For each class $C_m$, the Markov model has three parameters: the number of states (observation symbols) $K_m$, initial state probabilities $\Pi_m = \{\pi(v_i \mid C_m)\}$, and state transition probabilities $A_m = \{a(v_i, v_j \mid C_m)\}$, where

$$\pi(v_i \mid C_m) = P(O_1 = v_i \mid C_m) \quad (4)$$

$$a(v_i, v_j \mid C_m) = P(O_{t+1} = v_j \mid C_m \text{ and } O_t = v_i). \quad (5)$$

For learning the probabilities $\Pi_m$ and $A_m$, a new training set, $D_m = \{S^{(u)} \mid S^{(u)} \in C_m\}$, is formed generating $N$ sequences from each training image that belongs to the class $C_m$. Using this new training set, the probabilities are learned by maximum likelihood estimation that uses additive smoothing [27] with $\delta = 1$. The class likelihood is written as

$$P(S \mid C_m) = \pi(O_1 \mid C_m) \prod_{t=1}^{T-1} a(O_t, O_{t+1} \mid C_m). \quad (6)$$

The posteriors $P(C_m \mid S)$ are calculated by the Bayes rule assuming that each class is equally likely. The steps of the RMM to classify an unseen image are given in Fig. 4.

## III. Experiments

### A. Dataset

The dataset contains 3236 images of colon tissues of 258 randomly selected patients from the Pathology Department Archives in Hacettepe University Medical School. The tissues are stained with hematoxylin and eosin and the images are taken with a Nikon Coolscope Digital Microscope using a $20\times$ microscope objective lens and $480 \times 640$ image resolution.

We randomly divide the patients into two groups such that the training set contains 1644 images of the first half of the patients and the test set contains 1592 images of the remaining. We label each image with one of the three classes: normal, low-grade cancerous, or high-grade cancerous[2]. The training set contains 510 normal, 859 low-grade cancerous, and 275 high-grade cancerous tissues. The test set contains 491 normal, 844 low-grade cancerous, and 257 high-grade cancerous tissues.

### B. Comparisons

To investigate the effectiveness of our proposed method, we compare its results with those of the two sets of algorithms. The first set includes algorithms that define their features similar to the RMM but take different algorithmic steps for classification. We particularly implement these algorithms to understand the effectiveness of the sequence generation and Markov modeling steps proposed by the RMM. The second set includes algorithms that use different textural and structural features proposed by existing methods. We use them to compare the performance of the RMM and previous approaches.

*1) Algorithms with Similar Features:* First, we implement a grid-based counterpart of our method. In this ***GridBasedApproach***, an image is divided into grids, the same RMM features are extracted for the grids, and the grid features are averaged all over the tissue. Then, a support vector machine (SVM) with a linear kernel[3] is used for classification. This method directly uses grid features, as opposed to the RMM where grids are first discretized and then used for classification. Besides, it does not use resampling-based voting, which votes the decisions of a classifier obtained for the samples of the same image.

Second, we modify the previous grid-based approach so that it includes resampling-based voting. This ***VotingApproach*** generates $N$ samples from a test image similar to the RMM, classifies them using the learned SVM, and combines the decisions by majority voting. This method selects $T$ random grids to generate a sample and defines the features of the sample by averaging those of the selected grids.

[2]The images are labeled by Prof. C. Sokmensuer, MD, who is specialized in colorectal carcinomas.

[3]We also conduct our experiments using an RBF kernel. However, an RBF kerneled SVM is negatively affected from skewed class distribution and favors the low-grade over the high-grade cancerous class. Hence, we use a linear kerneled SVM, which is less likely to overfit the distribution of training data.

The previous two approaches directly use the extracted grid features, without discretizing the grids. The ***BagOfWordsApproach*** discretizes the grids into $K$ clusters in the same way of the RMM, forming the visual words of a vocabulary. Then, it divides a test image into grids, assigns each grid to its closest word, and uses the words' frequency to characterize the image. It also uses an SVM with a linear kernel for classification.

*2) Algorithms with Different Features:* First, we calculate the first-order histogram features. The ***IntensityHistogramFeatures*** include mean, standard deviation, kurtosis, and skewness values calculated on the intensity histogram of a gray-level tissue image [28]. To reduce the effects of noise or small intensity differences, pixel intensities are quantized into $N$ bins. Additionally, we calculate the grid-based version of these features. In calculating the ***IntensityHistogramGridFeatures***, instead of computing a single histogram for an entire image, we divide the image into grids, find the histogram of each grid, and average the features of the grids all over the image.

Next, we compute the ***CooccurrenceMatrixFeatures*** that use second-order statistics. They include energy, entropy, contrast, homogeneity, correlation, dissimilarity, inverse difference moment, and maximum probability features derived from a gray-level cooccurrence matrix of an entire image [14]. In our experiments, for a given distance, we compute cooccurrence matrices at eight different directions, $\theta = \{i\pi/4 \mid 0 \leq i \leq 7\}$, take their average to obtain a rotation invariant cooccurrence matrix, and calculate the features on this averaged matrix. Here gray-level pixel intensities are also quantized into $N$ bins. Likewise, as their grid-based version, we calculate the ***CooccurrenceMatrixGridFeatures***.

We use two sets of structural features in comparisons. The first set is extracted on color graphs [12]. In a color graph, nodes correspond to different types of tissue components located by a circle-fit algorithm, which has two parameters, and edges are defined by a Delaunay triangulation of these nodes. After coloring the edges according to their end nodes, colored versions of the average degree, average clustering coefficient, and diameter are defined as the ***ColorGraphFeatures***. The second set is extracted on a standard (colorless) Delaunay triangulation that is constructed on nuclear components located using the circle-fit algorithm. The ***DelaunayTriangulationFeatures*** include the average degree, average clustering coefficient, and diameter of the entire Delaunay triangulation as well as the average, standard deviation, minimum-to-maximum ratio, and disorder of edge lengths and triangle areas [8].

### C. Parameter Selection

The proposed resampling-based Markovian model (RMM) has four external model parameters: 1) the size of a window, in which the features of a sampled point are defined, 2) the number of states $K$ in a Markov model, 3) the length of a sequence $T$, and 4) the number of sequences $N$ generated for each image. Note that the number of states and observation symbols is the same in observable Markov models. In our experiments, we consider all possible combinations of the following parameter sets: $winSize = \{10, 20, 40, 80\}$, $K = \{4, 8, 16, 32, 64\}$, $T = \{10, 25, 50, 100, 150\}$, and $N = \{10, 25, 50, 100, 150\}$. Using 3-fold cross-validation on training images, we select the parameter combination that gives the maximum accuracy. The selected parameters are $winSize = 40$, $K = 64$, $T = 100$, and $N = 100$.

TABLE I
PARAMETERS OF THE ALGORITHMS TOGETHER WITH THEIR VALUES CONSIDERED IN CROSS VALIDATION

| | |
|---|---|
| *GridBasedApproach* | Grid size $= \{10, 20, 40, 80\}$ |
| *VotingApproach* | Grid size $= \{10, 20, 40, 80\}$<br>Number of grids $= \{10, 25, 50, 100, 150\}$<br>Trial number $= \{10, 25, 50, 100, 150\}$ |
| *BagOfWordsApproach* | Number of words $= \{4, 8, 16, 32, 64\}$<br>Grid size $= \{10, 20, 40, 80\}$ |
| *IntensityHistograms* | Bin number $= \{4, 8, 16, 32\}$ |
| *IntensityHistogramGrids* | Bin number $= \{4, 8, 16, 32\}$<br>Grid size $= \{10, 20, 40, 80\}$ |
| *CooccurrenceMatrices* | Bin number $= \{4, 8, 16, 32\}$<br>Distance $= \{5, 10, 20, 40\}$ |
| *CooccurrenceMatrixGrids* | Bin number $= \{4, 8, 16, 32\}$<br>Distance $= \{5, 10, 20, 40\}$<br>Grid size $= \{10, 20, 40, 80\}$ |
| *ColorGraphs* | Structuring element size $= \{3, 5, 7, 9\}$<br>Circle area threshold $= \{5, 10, \ldots, 50\}$ |
| *DelaunayTriangulations* | Structuring element size $= \{3, 5, 7, 9\}$<br>Circle area threshold $= \{5, 10, \ldots, 50\}$ |

The other algorithms have also parameters, which are listed in Table I. In addition to these, they have the SVM parameter $C$ as they use SVM classifiers with linear kernels [29]. Similarly, we use cross-validation on training images to select the parameters of each algorithm. The candidate values of each parameter are given in Table I. For all algorithms, the same set is considered for the SVM parameter: $C = \{1, 2, \ldots, 9, 10, 20, \ldots, 90, 100, 150, \ldots, 950, 1000\}$.

### D. Test Results

As tissue images typically contain a considerable amount of variance, classifiers usually require large amount of data to learn this variance better. However, acquiring large datasets from a large number of patients is quite difficult in this domain[4]. To address this problem, we conduct our experiments using all available training data as well as using less training data. For that, we randomly divide the training set into smaller subsets such that each subset includes $P$ % of the training data. For all algorithms, we repeat the experiments when $P$ is selected as 1, 2.5, 5, 10, 25, and 50 %. Since there are more than one subset for a selected $P$ value (e.g., 20 subsets when $P = 5$ %), we consider all subsets and report the average results. Besides, point selection in the RMM involves randomness. Thus, for the RMM, we repeat the experiments for 40 times with the selected parameters and also consider these runs in average computation. Fig. 5 plots the overall test set accuracies as a function of $P$. Additionally, Table II reports the class accuracies[5] for the selected $P$ values.

[4]For the first sight, our dataset seems to be a counter example. However, it is worth noting that the preparation of this dataset, which includes case selection, archive search, slide examination, image acquisition, and labeling steps, takes more than three years. Thus, this dataset is actually a good example that indicates the difficulty of acquiring large datasets in this domain.

[5]For a particular class, the class accuracy is calculated considering only the results of the classifier obtained on the images of this particular class.

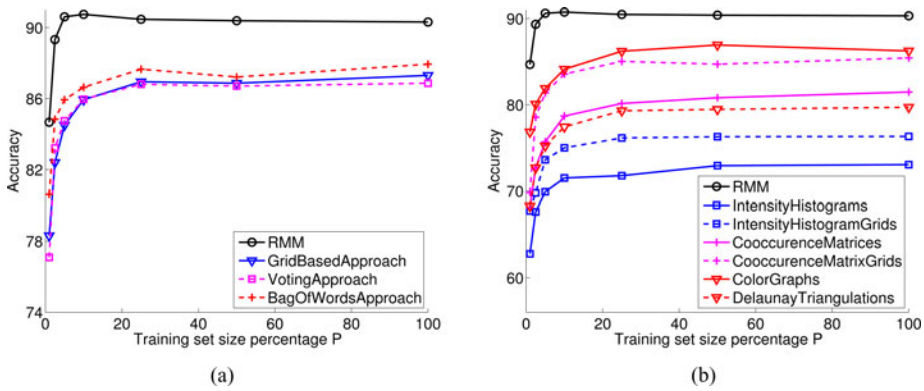


Fig. 5. Performance of the algorithms as a function of the training set size: (a) The test set accuracies of the algorithms that use features similar to those of the RMM and (b) the test set accuracies of the algorithms that use features different than those of the RMM.

TABLE II
CLASSIFICATION ACCURACIES ON THE TEST SET AND THEIR STANDARD DEVIATIONS

| | | $P = 100\%$ | | | $P = 10\%$ | | | $P = 5\%$ | | | $P = 1\%$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Normal | Low | High | Normal | Low | High | Normal | Low | High | Normal | Low | High |
| **Similar Features** | RMM | 95.64 (±0.18) | 87.77 (±0.32) | 88.56 (±0.39) | 95.22 (±0.58) | 89.45 (±1.99) | 86.46 (±2.94) | 94.69 (±1.37) | 90.76 (±2.84) | 82.32 (±5.27) | 86.46 (±6.56) | 95.67 (±2.92) | 45.12 (±19.42) |
| | *GridBasedApproach* | 91.65 | 85.31 | 85.60 | 90.31 (±3.00) | 84.30 (±3.28) | 82.96 (±3.13) | 87.25 (±5.39) | 85.12 (±4.97) | 77.20 (±8.15) | 84.67 (±9.38) | 83.56 (±9.60) | 48.96 (±18.97) |
| | *VotingApproach* | 90.84 | 84.72 | 86.38 | 90.53 (±2.95) | 84.02 (±3.21) | 83.62 (±2.96) | 87.65 (±5.59) | 84.84 (±4.73) | 78.89 (±7.35) | 82.31 (±9.38) | 84.25 (±9.60) | 43.61 (±18.97) |
| | *BagOfWordsApproach* | 94.91 | 87.32 | 76.65 | 93.73 (±1.88) | 90.20 (±2.04) | 61.40 (±9.18) | 92.31 (±2.84) | 90.59 (±4.05) | 58.48 (±9.25) | 88.62 (±5.43) | 87.30 (±6.89) | 43.31 (±18.88) |
| **Different Features** | *IntensityHistograms* | 80.65 | 69.55 | 70.04 | 79.04 (±3.58) | 69.64 (±5.97) | 63.23 (±8.84) | 77.08 (±4.84) | 69.54 (±9.44) | 57.43 (±10.19) | 64.69 (±13.78) | 71.84 (±12.93) | 29.39 (±14.38) |
| | *IntensityHistogramGrids* | 78.82 | 74.17 | 78.60 | 77.70 (±3.33) | 73.35 (±5.61) | 75.25 (±3.80) | 75.67 (±6.21) | 73.35 (±7.65) | 70.56 (±6.43) | 69.22 (±13.85) | 75.90 (±12.96) | 38.18 (±18.57) |
| | *CooccurrenceMatrices* | 83.10 | 81.64 | 77.82 | 78.88 (±4.09) | 81.15 (±4.02) | 70.23 (±8.32) | 75.17 (±6.75) | 79.26 (±6.37) | 65.04 (±11.12) | 66.66 (±15.27) | 76.17 (±10.58) | 44.54 (±18.49) |
| | *CooccurrenceMatrixGrids* | 87.58 | 84.12 | 85.60 | 83.77 (±3.08) | 83.95 (±3.05) | 81.87 (±4.89) | 81.45 (±5.01) | 82.64 (±5.39) | 77.02 (±9.44) | 71.97 (±12.92) | 75.13 (±11.15) | 48.26 (±20.50) |
| | *ColorGraphs* | 92.67 | 82.46 | 86.38 | 88.37 (±3.35) | 84.23 (±2.55) | 75.64 (±5.37) | 85.22 (±4.38) | 85.79 (±5.26) | 62.70 (±8.62) | 80.89 (±9.32) | 82.93 (±7.76) | 49.03 (±20.92) |
| | *DelaunayTriangulations* | 89.61 | 71.56 | 87.55 | 86.80 (±1.91) | 72.75 (±7.63) | 74.94 (±8.62) | 82.03 (±6.20) | 75.31 (±8.68) | 61.93 (±9.13) | 72.96 (±12.46) | 73.60 (±10.24) | 41.81 (±17.11) |

The results are obtained when all training data are used (when $P = 100\%$ and when limited training data are used (when $P = 10\ \%$, $P = 5\%$ and $P = 1\%$.

When all training data are used ($P = 100\ \%$), the results show that the RMM improves the accuracy of the other algorithms; the McNemar's test with Bonferroni correction gives that the overall accuracy improvement is statistically significant with $\alpha = 0.05$. This may be due to the following: A tissue image typically contains irrelevant information and noise at the pixel level. Thus, feature extraction, which transforms the image into a feature domain, may result in important data loss. Since the RMM generates sequences (features) of the same image using different image subregions, which can be very divergent from one sequence to another, the sequences are expected to include different data loss. This is opposed to the case of many algorithms that extract just a single feature vector from the same image. In that sense, the RMM contributes more information to the feature domain, although it does not add any new information in the image (entire data) domain.

The results also reveal that the algorithms that use grid-based aggregation usually perform better than those that use the image in its entirety. This is attributed to the issue of finding a constant texture for an image that contains irrelevant regions in the context of classification (see Fig. 2). The RMM, which can be considered as an aggregation method, further improves these grid-based algorithms. The RMM yields better accuracies than the *GridBasedApproach* and the *VotingApproach*, which do not use the discretized grids in their classification. This indicates the usefulness of state definition of the RMM. Besides, comparing the RMM against the *VotingApproach*, the results show that generating sequences is more effective in resampling-based voting. The *BagOfWordsApproach* uses state definition but does not employ resampling-based voting in its classification. The RMM improves the performance of the *BagOfWordsApproach*, showing the effectiveness of using resampling-based voting. This improvement is especially observed for correct classification of high-grade cancerous tissues; as future research work, one could incorporate the proposed framework into a bag-of-words approach. Additionally, none of the algorithms represent an image using sequences. The results also indicate the importance of this representation.

When partial data are used for learning, we observe that the test set accuracies decrease with the decrease in the number of training samples. For the other methods, this decrease becomes noticeable when $P \leq 25\ \%$ (i.e., when $\leq 411$ samples are used for training). However, the proposed RMM is able to keep the test accuracy high even when 5 % of the training data are used.

Note that, in these plots, there is a slight increase in the accuracy of the RMM when $P$ decreases. This is due to the unbalanced class distribution in the test set. As $P$ decreases, the accuracy of the low-grade class increases at the expense of decreasing the high-grade class accuracy. As the number of low-grade cancerous tissue images is relatively high, this slightly increases the overall accuracy.

The high performance of the RMM is attributed to the following. The other algorithms do not attempt to vary training images for better generalizations. They just use the available training images in their current form. On the other hand, the RMM has the flexibility to increase the variety of training images by resampling. It can adapt itself to the cases where there are less training images by increasing the number of sequences it generates from an individual image. In the experiments, we use this property and adjust the number of generated sequences according to the value of $P$ (e.g., if $N$ sequences are generated when the entire dataset is used, $20 \times N$ sequences are generated when $P = 5$ %). This property becomes especially important when the training set becomes smaller. This may be one of the major reasons behind obtaining stable accuracy results until $P = 5$ %. When $P < 5$ %, a decrease is observed also for the RMM. This is due to a relatively higher accuracy decrease in high-grade cancerous tissues (Table II). The number of high-grade cancerous tissue images is relatively smaller in the training set and resampling is not able to sufficiently vary the data with such a small size.

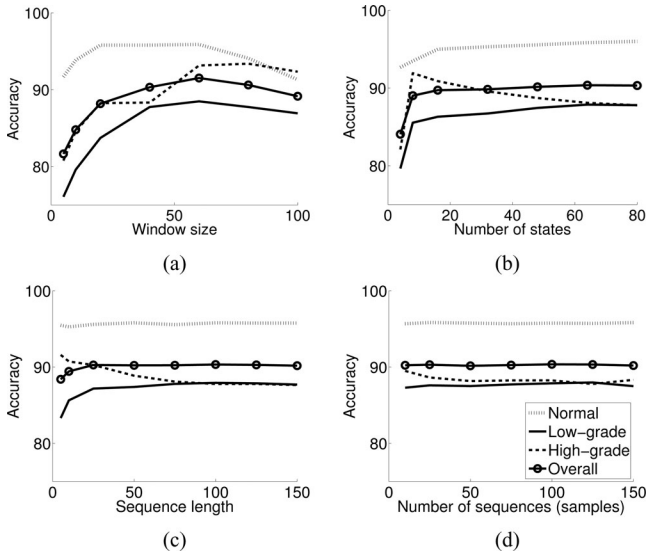


Fig. 6. Test accuracies as a function of the model parameters: (a) Window size, (b) number of states, (c) sequence length, and (d) number of sequences.

TABLE III
TEST SET RESULTS FOR ALTERNATIVE DESIGN CHOICES

| | Overall | Difference |
|---|---|---|
| Proposed RMM | 90.32 ± 0.18 | – |
| RMM perturbing entire images | 87.98 ± 0.20 | Yes |
| RMM using SIFT points | 90.75 ± 0.14 | No |
| RMM with random initial points | 90.28 ± 0.15 | No |
| RMM using zero-order Markov models | 87.54 ± 0.16 | Yes |

### *E. Parameter analysis*

The RMM has four external parameters: window size, number of states, sequence length, and number of sequences. The effects of each parameter on test accuracies are investigated. For that, three of the four parameters are fixed and the accuracy is observed as a function of the other parameter. Using the entire training data for learning, we give the parameter analysis performed on the test set in Fig. 6.

The window size controls the size of a region, in which the features of a single data point are defined. Smaller regions do not cover enough pixels to characterize the data points satisfactorily, resulting in lower accuracies. On the other hand, larger regions cover pixels of different characteristics, and hence, give too generic features for the data points. This slightly decreases the classification accuracy.

The number of states determines the number of observation symbols in an observable Markov model. In the RMM, observation symbols represent tissue subregions with different characteristics. Thus, larger values of this parameter allow increasing the variety of subregions. This is effective in increasing the accuracy. On the other hand, larger numbers also increase the number of transition probabilities to be estimated. If this estimation is not good enough, larger numbers may decrease the accuracy. Although this effect is not seen in Fig. 6(b), we observe it when we use less data (smaller $P$) for estimation. In such cases, better accuracies could be obtained by using smaller values of this parameter.

The sequence length affects the size of a region a sample covers. If it is selected too small, the sample does not cover large enough area to characterize the image. Increasing the length increases the accuracy.

The number of sequences controls the number of samples generated to represent a tissue image. If it is selected too small, there is a risk of not obtaining representative samples from the image. Moreover, the RMM does not use any normalization to characterize its windows. Hence, it may label two biologically similar windows differently (e.g., a window comprising a small luminal region and another one comprising a large luminal region can be labeled differently). On the other hand, this may be offset by the sequence generation step since the RMM is capable of generating a variety of sequences for the same image, provided that a sufficient number of sequences is generated. Thus, the number of sequences should not be selected too small. Additionally, it should be more than one to use the voting scheme in classification.

In addition to these parameters, the RMM includes implicit design choices. In order to understand their effects, we repeat our experiments using the same external parameters but with alternative choices. We summarize our results in Table III. In this table, we report the overall accuracies as well as whether or not there exists statistically significant difference between our design choices and their alternatives (with $\alpha = 0.05$).

The RMM perturbs images by taking their different parts; however, one may prefer perturbing the entire images. For that, an image can be divided into windows and these windows can

be characterized with states and reordered randomly. Our experiments reveal that perturbing entire images is significantly less effective. We attribute this to the diversity of the generated sequences. When an entire image is used, the diversity is expected to be smaller since all sequences contain the same set of windows. On the other hand, the RMM generates sequences that contain different windows, which is expected to increase the diversity among the generated sequences. We also compare the diversity quantitatively by considering images one by one, measuring the variation in the sequences of each image, and taking the sum of the variation over all images. For a given image, the variation is measured by calculating a transition probability matrix for each of its individual sequences and computing the variance of probabilities that belong to the same transition. This variance indicates the degree of how the frequency of a particular transition (from one state to another) varies in different sequences of the same image. Then these variances are summed over all transitions. The results obtained on the training images show that the RMM increases the variance sum from 7.67 to 16.22, compared to its counterpart.

In order to select its points, the RMM follows a random approach. We repeat our experiments selecting them among the SIFT points [30]. The results show that the use of the SIFT points gives similar results. This indicates that compared to the random ones, the SIFT points do not carry additional information for this particular application. However, one may work on defining domain specific salient points and use them in selection. This can be considered as future work.

In sequence generation, the RMM orders the points starting from the one closest to the top-left corner. However, one may select the initial point randomly. Our experiments show that this yields similar results. This may be due to the following: First, as the RMM employs the same greedy method, the sequences generated for the same points will contain lots of similar subsequences although their initial points are different. Second, it uses many sequences instead of a single one. Some of the sequences can be similar to those of the other images since the same initial point selection is used for all images.

To learn class probabilities, the RMM uses first order Markov models. We explore the effects of using zero-order Markov models, which assume no dependency between the subsequent states. This use gives significantly worse results. Here, it is also possible to use higher-order Markov models. Nevertheless, it requires learning more number of parameters (transition probabilities). This, however, may decrease the accuracy if there are not sufficient occurrences of successive states in training samples. This may especially become a problem when there are limited training samples.

## IV. Conclusion

This paper successfully addresses the issue of having limited labeled training data in the domain of histopathological tissue image classification. To this end, it presents a new resampling framework that generates multiple sequences from an image and models them using first order discrete Markov processes.

The proposed resampling-based Markovian model (RMM) is tested on 3236 colon tissue images. The experiments demonstrate that the proposed RMM is more effective to keep the accuracy high when less training data are used for learning. This is attributed to the ability of the RMM to increase the generalization capacity of a learner by increasing the size and variation of the training data. Additionally, the experiments show that the voting scheme, which combines the decisions of its sequences to classify an image, is also effective in increasing the classification accuracy.

As noted earlier, the proposed RMM does not impose any particular feature type to characterize data points. In this work, we use a set of simple features since they are easy to extract and do not introduce an additional parameter, unlike those used in comparisons (e.g., the *ColorGraphs* approach involves two additional parameters). One future research direction is to focus on feature extraction and incorporate different features in the proposed framework. For instance, one can use textural features for a selected data point by centering a window at this point and defining the texture of pixels located in this window. It is also possible to extract structural features by defining a graph on the tissue and calculating local features for the graph nodes. In this case, data point selection should be restricted so that only the node centroids are selected and the local features are used to characterize the selected points.

The RMM uses Markov modeling since it is known as one of the simplest and most effective ways for modeling sequences. However, one may explore the use of other sequence modeling methods such as hidden Markov models and recurrent neural networks. Additionally, instead of using sequences, a feature vector can be defined for an image using the features of its selected points and such feature vectors can be used by different classifiers such as SVMs.

Although it is particularly designed for histopathological images and the experiments are conducted on colon tissues, the proposed method has a potential to be used for different types of images as well as different types of tissues. This can also be considered as a future research direction of the paper.

## Appendix

We provide the pseudocode of observation symbol learning and sequence generation in Algorithms 1 and 2, respectively.

**Algorithm 1** LearnObservationSymbols

**Input:** training set $\tau$, window size $winSize$, number of observation symbols $K$
**Output:** observation symbols $V$

$\Phi \leftarrow \emptyset$
**for** $i = 1 \rightarrow |\tau|$ **do**
  **for** $j = 1 \rightarrow 100$ **do**
    $P \leftarrow$ SelectRandomPoint$(\tau_i)$
    $F \leftarrow$ ExtractFeatures$(\tau_i, P, winSize)$
    $\Phi \leftarrow \Phi \cup \{F\}$
  **end for**
**end for**
$V \leftarrow$ KmeansClustering$(\Phi, K)$

**Algorithm 2** SEQUENCEGENERATION

**Input:** image $I$, observation symbols $V$, window size $winSize$, sequence length $T$

**Output:** sequence $S$

1: $\mathcal{P} \leftarrow \emptyset, \quad \mathcal{O} \leftarrow \emptyset$
2: **for** $t = 1 \rightarrow T$ **do**
3: $P_t \leftarrow$ SELECTRANDOMPOINT$(I)$
4: $F_t \leftarrow$ EXTRACTFEATURES$(I, P_t, winSize)$
5: $O_t \leftarrow$ ASSIGNOBSERVATIONSYMBOL$(F_t, V)$
6: $\mathcal{P} \leftarrow \mathcal{P} \cup \{P_t\}, \quad \mathcal{O} \leftarrow \mathcal{O} \cup \{O_t\}$
7: **end for**
8: $S \leftarrow$ ORDERPOINTS$(\mathcal{P}, \mathcal{O})$

## REFERENCES


[1] A. Jemal, R. Siegel, J. Xu, and E. Ward, "Cancer statistics 2010," *CA-Cancer J. Clin.*, vol. 60, no. 5, pp. 277–300, 2010.

[2] A. Andrion, C. Magnani, P. G. Betta, A. Donna, F. Mollo, M. Scelsi, P. Bernardi, M. Botta, and B. Terracini, "Malignant mesothelioma of the pleura: Inter-observer variability," *J. Clin. Pathol.*, vol. 48, no. 9, pp. 856–860, 1995.

[3] W. Wang, J. A. Ozolek, and G. K. Rohde, "Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images," *Cytometry Part A*, vol. 77A, no. 5, pp. 485–494, 2010.

[4] H.-K. Choi, T. Jarkrans, E. Bengtsson, J. Vasko, K. Wester, P.-U. Malmstrom, and C. Busch, "Image analysis based grading of bladder carcinoma. Comparison of object, texture and graph based methods and their reproducibility," *Anal. Cell. Pathol.*, vol. 15, pp. 1–18, 1997.

[5] J. Gil, H. Wu, and B. Y. Wang, "Image analysis and morphometry in the diagnosis of breast cancer," *Microsc. Res. Techniq.*, vol. 59, pp. 109–118, 2002.

[6] A. N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, G. Bhanot, and A. Madabhushi, "Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 3, pp. 642–653, 2010.

[7] B. Weyn, G. van de Wouwer, S. Kumar-Singh, A. Van Daele, P. Scheunders, E. van Marck, and W. Jacob, "Computer-assisted differential diagnosis of malignant mesothelioma based on syntactic structure analysis," *Cytometry*, vol. 35, pp. 23–29, 1999.

[8] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *Proc. 5th IEEE Int. Symp. Biomed Imaging: From Nano to Macro*, Paris, May. 14–17, 2008, pp. 496–499.

[9] C. Demir, S. H. Gultekin, and B. Yener, "Learning the topological properties of brain tumors," *IEEE ACM T. Comput. Bi.*, vol. 2, no. 3, pp. 262–270, Jul./Sep. 2005.

[10] C. Gunduz-Demir, "Mathematical modeling of the malignancy of cancer using graph evolution," *Math. Biosci.*, vol. 209, no. 2, pp. 514–527, 2007.

[11] C. Demir, S. H. Gultekin, and B. Yener, "Augmented cell-graphs for automated cancer diagnosis," *Bioinformatics*, vol. 21, no. Suppl 2, pp. ii7–ii12, 2005.

[12] D. Altunbay, C. Cigir, C. Sokmensuer, and C. Gunduz-Demir, "Color graphs for automated cancer diagnosis and grading," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 3, pp. 665–674, 2010.

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

[14] A. N. Esgiar, R. N. G. Naguib, B. S Sharif., M.K. Bennett, and A. Murray, "Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa," *IEEE T. Inf. Technol. Biomed.*, vol. 2, no. 3, pp. 197–203, Sep. 1998.

[15] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "A boosted Bayesian multi-resolution classifier for prostate cancer detection from digitized needle biopsies," *IEEE Trans. Biomed. Eng.*, 2011, in press. DOI: 10.1109/TBME.2010.2053540.

[16] B. Weyn, G. van de Wouver, M. Koprowski, A. van Daele, K. Dhaene, P. Scheunders, W. Jacob, and E. van Marck, "Value of morphometry, texture analysis, densitometry, and histometry in the differential diagnosis and prognosis of malignant mesothelia," *J. Pathol.*, vol. 189, pp. 581–589, 1999.

[17] K. Jafari-Khouzani and H. Soltanian-Zadeh, "Multiwavelet grading of pathological images of prostate," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 6, pp. 697–704, 2003.

[18] O. Sertel, J. Kong, H. Shimada, U. V. Catalyurek, J. H. Saltz, and M. N. Gurcan, "Computer-aided prognosis of neuroblastoma on whole slide images: Classification of stromal development," *Pattern Recognit.*, vol. 42, no. 6, pp. 1093–1103, 2009.

[19] H. Qureshi, O. Sertel, N. Rajpoot, R. Wilson, and M. N. Gurcan, "Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification," in *Proc. 11th Int. Conf, Medical Image Computing and Computer Assisted Intervention*, pp. 196–204, 2008.

[20] P.-W. Huang and C.-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE Trans. Med. Imaging*, vol. 28, no. 7, pp. 1037–1050, 2009.

[21] L. E. George and K. H. Sager, "Breast cancer diagnosis using multi-fractal dimension spectra," in *Proc. IEEE Int. Conf. Signal Process. Commun.*, 2007, pp. 592–595.

[22] O. R. Duda, E. P. Hart, and G. D. Stork, *Pattern Classification*. New York: Wiley Interscience, 2001.

[23] A. Ozcift, "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis," *Comput. Biol. Med.*, vol. 41, no. 5, pp. 265–271, 2011.

[24] U. Moller, "Resampling methods for unsupervised learning from sample data," in *Machine Learning*, A. Mellouk and A. Chebira, Eds. Cape Town, SA: InTech, 2009, pp. 289–304.

[25] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 8, pp. 800–810, Aug. 2001.

[26] H. Yu and E. R. Hancock, "String kernels for matching seriated graphs," in *Proc. Int. Conf. Pattern Recog.*, Hong-Kong, 2006, pp. 224–228.

[27] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Comput. Speech Lang.*, vol. 13, p. 359, 1999.

[28] M. Wiltgen, A. Gerger, and J. Smolle, "Tissue counter analysis of benign common nevi and malignant melanoma," *Int. J. Med. Inform.*, vol. 69, pp. 17–28, 2003.

[29] C.-C. Chang and C.-J. Lin. (2001). "LIBSVM: A library for support vector machines," [Online]. Available:http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

[30] A. Vedaldi and B. Fulkerson. (2008). "VLFeat: An open and portable library of computer vision algorithms," [Online]. Available:http://www.vlfeat.org.


**Erdem Ozdemir** received the B.S. and M.S. degrees in computer engineering from Bilkent University, Turkey, in 2008 and 2011, respectively. He is currently a Ph.D. student under the supervision of Dr. Gunduz-Demir in the Department of Computer Engineering at Bilkent University.

His research interests include the use of structural representations for classification and retrieval of histopathological images.

**Cenk Sokmensuer** received the medical degree and pathology training from Hacettepe University School of Medicine, Turkey.

He is currently a Professor of pathology at Hacettepe University, Turkey. As a visiting scholar, he worked in Harvard University in the USA during 2003–2004, in Necker Children Hospital in France in 1998, and in Victor Dupuy Hospital in France in 1992. His specialization includes pathology of gastrointestinal system, liver, and endocrine system.

**Cigdem Gunduz-Demir** (S'05–M'06) received the B.S. and M.S. degrees in computer engineering from Bogazici University, Turkey, in 1999 and 2001, respectively, and the Ph.D. degree in computer science from Rensselaer Polytechnic Institute, New York, in 2005.

She is currently an Assistant Professor with the Department of Computer Engineering at Bilkent University. Her research interests include development of new biocomputational models and application of pattern recognition, and computer vision algorithms for medical image analysis.