RESAMPLING-BASED MARKOVIAN MODELING FOR AUTOMATED CANCER DIAGNOSIS

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING

AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Erdem Özdemir

August, 2011

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Çiğdem Gündüz Demir (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Cevdet Aykanat

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Fatoş Yarman Vural

Approved for the Graduate School of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Graduate School

ABSTRACT

RESAMPLING-BASED MARKOVIAN MODELING FOR AUTOMATED CANCER DIAGNOSIS

Erdem Özdemir
M.S. in Computer Engineering
Supervisor: Assist. Prof. Dr. Çiğdem Gündüz Demir
August, 2011

Correct diagnosis and grading of cancer is very crucial for planning an effective treatment. However, cancer diagnosis on biopsy images involves visual interpretation of a pathologist, which is highly subjective. This subjectivity may, however, lead to selecting suboptimal treatment plans. In order to circumvent this problem, it has been proposed to use automatic diagnosis and grading systems that help decrease the subjectivity levels by providing quantitative measures. However, one major challenge for designing these systems is the existence of high variance observed in the biopsy images due to the nature of biopsies. Thus, for successful classifications of unseen images, these systems should be trained with a large number of labeled images. However, most of the training sets in this domain have limited size of labeled data since it is quite difficult to collect and label histopathological images. In this thesis, we successfully address this issue by presenting a new resampling framework. This framework relies on increasing the generalization capacity of a classifier by augmenting the size and variation in the training set. To this end, we generate multiple sequences from an image, each of which corresponds to a perturbed sample of the image. Each perturbed sample characterizes different parts of the image, and hence, they are slightly different from each other. The use of these perturbed samples for representing the image increases the size and variability of the training set. These samples are modeled with Markov processes which are used to classify unseen image. Working with histopathological tissue images, our experiments demonstrate that the proposed framework is more effective for both larger and smaller training sets compared against other approaches. Additionally, they show that the use of perturbed samples is effective in a voting scheme which boosts the performance of the classifier.

Keywords: Histopathological image analysis, automated cancer diagnosis, resampling, Markov models, cancer.

ÖZET

OTOMATİK KANSER TANISI İÇİN TEKRAR ÖRNEKLEME BAZLI MARKOV MODELLEME

Erdem Özdemir
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Yar. Doç. Dr. Çiğdem Gündüz Demir
August, 2011

Doğru kanser tanısı ve derecelendirilmesi, etkili bir tedavi planı için önemlidir. Ancak, biyopsi görüntüleri üzerinde yapılan kanser tanısı, patologların görsel olarak yorumlamasına dayanır, bu ise öznellik taşır. Bu öznellik, etkili olmayan tedavi planlarının uygulanmasına yol açabilir. Tanıdaki öznelliği azaltmak amacıyla, ölçülebilir değerler üzerinden otomatik kanser tanısı ve derecelendirmesi yapan sistemler önerilmiştir. Biyopsi görüntülerinde varolan değişim bu tür sistemlerin tasarlanmasında en büyük sorunlardan birini oluşturur. Bu tür sistemlerin biyopsi görüntülerini doğru sınıflandırabilmesi için fazla sayıda öğrenme örneği ile eğitilmesi gerekir. Ancak, histopatolojik görüntü alanındaki öğrenme kümeleri, görüntü toplama ve bu görüntüleri etiketlemedeki zorluklardan ötürü genelde az sayıda örnek içerir. Biz bu çalışmamızda bu probleme karşı, öğrenme kümesindeki örnek sayısını ve varyansını artırarak sınıflandırıcının genelleme kapasitesini artıran, yeni bir tekrar örnekleme yöntemi sunmaktayız. Bunu yapabilmek için, görüntü üzerinden her biri, görüntünün değiştirilmiş örneğine denk gelen diziler oluşturulur. Bu değiştirilmiş örneklerin her biri, görüntü üzerinde değişik alt bölgeleri nitelendirdirir ve dolayısıyla birbirinden farklıdır. Bu örneklerin öğrenmede kullanılması ise öğrenme kümesinin büyüklüğünü ve varyansını artırır. Markov modeller ile bu örnekler modellenir ve etiketlenmemiş örneklerin sınıflandırılmasında kullanılır. Histopatolojik görüntüler üzerinde yapılan testlerde, sunulan bu yöntemin hem büyük hem de küçük boyutlu öğrenme kümelerinde diğer yöntemlere göre daha başarılı olduğu görülmektedir. Ayrıca değiştirilmiş örneklerin oylama yönteminde kullanılması sınıflandırıcının performansını artırmaktadır.

Anahtar sözcükler: Histopatolojik görüntü analizi, kanser tanı ve derecelendirilmesi, tekrar örnekleme, Markov modelleri, kanser.

Acknowledgement

I would like to thank my advisor Assist. Prof. Dr. Çiğdem Gündüz Demir for her endless support and guidance throughout this thesis. Without her support I would not finish this thesis. I am also grateful to my jury members Prof. Dr. Cevdet Aykanat and Prof. Dr. Fatoş Yarman Vural for their time reading and evaluating this thesis.

I owe my deepest gratitude to my family, Ali, Kezban Özdemir and my sisters Safiye Nur and Nurdan and my fiance Nurcan. This thesis would not be possible without their help, patience and support.

I thank my friends in the office Salim, Shatlik, Onur, Alper, Barış and Burak for their friendship, support and for amusing times in the office. I am also grateful to Salim for his help in drawing some figures. Last, but by no means least, I would like to thank everyone in the Gunduz group: Can, Barış, Burak, Salim, Çağrı and Gülden.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

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 [35]. Although there are many screening techniques such as colonoscopy, sigmoidoscopy, and stool test, its final diagnosis and grading are based on histopathological assessment of biopsy tissue samples. In this assessment, pathologists decide 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, which leads to interandintra observer variability especially in grading [3, 43]. This variability may result in suboptimal treatment of the disease [13]. Thus, it has been proposed to use computational methods. These methods would help pathologists make more objective assessment by providing quantitative measures.

1.1 Motivation

The previous methods provide automated classification systems that use a set of features to model the difference between the normal tissue appearance and

the corresponding abnormalities. These features are usually defined by the motivation of mimicking a pathologist, who uses morphological changes in cells 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 cells. These features can be used to characterize an individual cell [57, 65] as well as an entire tissue by aggregating the features of its cells [60, 10]. Extraction of morphological features requires determining the exact locations of cells beforehand, which is, however, very challenging for histopathological tissue images due to their complex nature [28].

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. These studies use different graph generation methods including Delaunay triangulations (and their dual Voronoi diagrams) [5, 68, 58], minimum spanning trees [10, 17], probabilistic graphs [14, 30], and weighted graphs [15]. To model topological tissue changes better, Dogan *et al.* 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 [2]. The main challenge of defining structural features is the difficulty of locating the tissue components. The incorrect localization may affect the success of the structural methods.

Textural methods avoid difficulties relating to correct localization of cells (and other components) 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 [59], co-occurrence matrices [21, 18], run-length matrices [67], multiwavelet coefficients [56], local binary patterns [54, 51], and fractal geometry [59, 32]. Textural methods generally make pixel level analysis, hence, they may negatively be affected by the noise in the intensity levels of the pixels. Moreover, textural

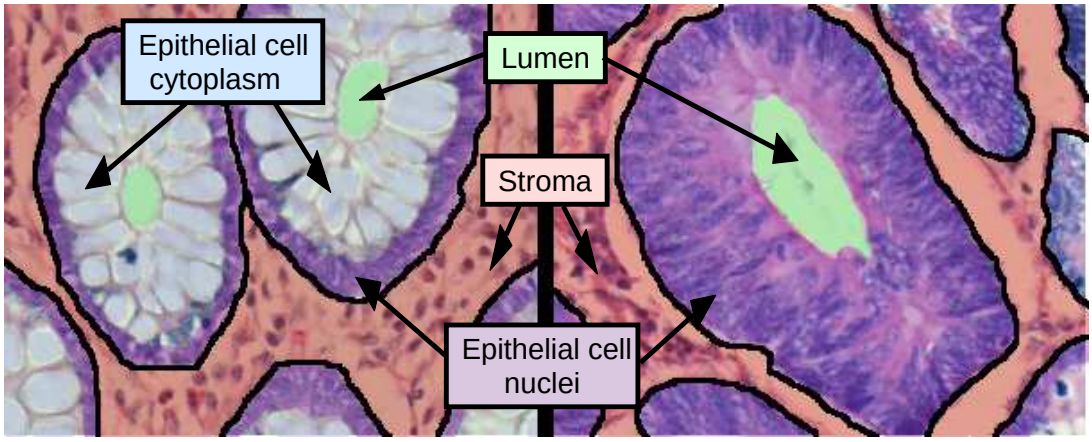


Figure 1.1: Cytological components in normal and cancerous colon tissues. Different components are illustrated with different colors: green for luminal regions, red for stromal regions, purple for epithelial cell nuclei, and blue for epithelial cell cytoplasms. Colon glands are confined with black boundaries.

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 [21]. Although, grid-based approaches usually improve accuracies, they may still have difficulties arising from the existence of irrelevant regions in tissues. For example, for diagnosis and grading of colon adenocarcinoma, which accounts for 90-95 percent of all colorectal cancers, pathologists examine glandular tissue regions since this cancer type originates from glandular epithelial cells and causes deformations in glands (Figure. 1.1). Non-glandular regions, which do not include epithelial cells, are irrelevant within the context of colon adenocarcinoma diagnosis. Moreover, such non-glandular regions can be of different sizes (Figure 1.2). Thus, directly including these regions into texture computation may give unstable classifications, resulting in lower accuracies [26]. Aggregation methods that consider the existence of such irrelevant regions have potential to give better accuracies.

Additionally, all classification algorithms face a common difficulty regardless of their feature types: limited training data to be generalized to unseen cases. This problem exists in various domains such as classification of data streams [42],

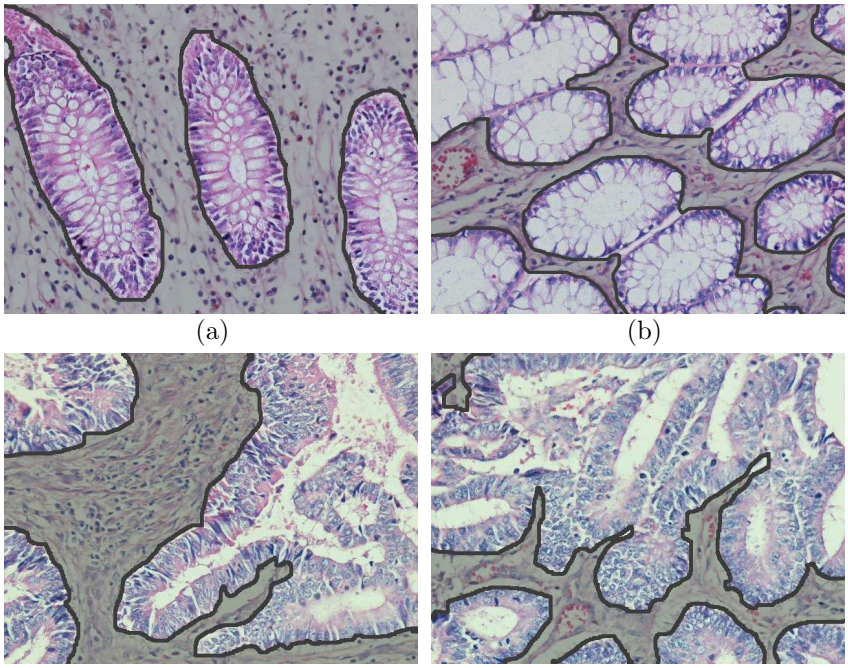


Figure 1.2: Histopathological images of colon tissues: (a)-(b) normal and (c)-(d) cancerous. Non-glandular regions in images are shaded with gray.

remote sensing [33, 27], speech recognition [38], information theory [71], and biomedical engineering [46, 9, 62, 4]. In our domain, due to the nature of the system, there exists large variance even among tissues, even the tissues of the same class. This is mainly due to irregularity of tumor growth [32]. The variance becomes even larger due to nonideal steps followed in tissue preparation as well as differences in tissue preparation and image acquisition steps. This problem has been stated in [62] as need of large datasets for robust applications in computer aided diagnosis systems. Thus, in order for a classification system to make successful generalizations, it 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 perturbations in tissue images, also leading to unstable classifications. There have been studies in active learning to address the issue of labeling cost; the studies have proposed to reduce the number of the required samples that are to be labeled [70, 40]. However,

active learning is a selection approach of the samples to be labeled in a large dataset and does not help learn when there are only limited data available. Simple resampling techniques have been used to resolve unbalanced data problem by increasing the number of samples in classes that include relatively less number of samples [49]. However, simple resampling techniques do not increase the variety since it dublicates the samples in the original dataset without introducing any modifications.

1.2 Contribution

In this thesis, 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, the main contributions are the introduction of a new resampling method to simulate the perturbations 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 working in this domain, 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 perturbations 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 perturbations with Markov processes provides an effective method in modeling the irrelevant regions.

1.3 Outline of the Thesis

This thesis is structured as follows. In Chapter 2, we give background information about the problem domain and summarize the existing approaches from literature. In Chapter 3, we present the details of our method including how perturbed samples from an image are generated and how these perturbed samples are used in learning and classification by modeling them with Markov chains. In Chapter 4, we explain the dataset, the test environment, and comparison methods. Then, in the same chapter, we report the test results and give a discussion of the results. Finally, we finalize the thesis with a conclusion and its future aspects, in Chapter 5.

Chapter 2

Background

In this chapter, we first present domain description in which we explain a specific cancer type colon adenocarcinoma and how colon tissues undergo deformation as a consequence of adenocarcinoma. Then we explain the classes that a tissue image can be classified to and how they are different from each other. Next, we present a summary of textural, morphological, and structural approaches in the literature for automated cancer diagnosis. Finally, we discuss the problem of having limited training dataset and discuss active learning, semi-supervised learning and resampling techniques.

2.1 Domain Description

In this thesis, we focus on colon adenocarcinoma, which is estimated to be responsible of 90-95 percent of all types of colorectal cancer. Colorectal cancer is the third most common cancer type among men and women in the USA [35]. Colon adenocarcinoma affects glandular tissue in the inner wall of the colon, which is responsible for secreting materials to lubricate waste products and absorbing water and some minerals from waste products before excretory. Colon adenocarcinoma starts at inner wall of the colon then spreads to the entire colon, potentially to the lymphatic system and the other organs as well. If it spreads, it may be fatal.

However, colon adenocarcinoma is one of the most curable cancer types if it is early detected. Screening tests such as colonoscopy and flexible sigmoidoscopy help early detection of colon adenocarcinoma without need of the surgery and according to [35] there is an observed decrease of colon adenocarcinoma due to the increased prevalence of these tests. Although these screening tests are important at early detection of colon adenocarcinoma. The final diagnosis together with its grade, can be made after examining biopsy sample by a pathologist under a microscope. For that, a small amount of the tissue of the concerned area is removed from the human body, and then this removed tissue is cut into thin slices. These thin slices are named as sections and the process is known as sectioning. Subsequently, for better visualization of these sections under a microscope, they are stained with a chemical process, which is named as staining. Staining gives contrast to the tissue, highlighting its special components for better visualization. The routinely used staining technique is hematoxylin and eosin. Hematoxylin stains nucleic acids deep blue and eosin stains cytoplasm pink as a result of a chemical reaction [24]. An illustrative example of an histopathological image from a colon tissue is given in Figure 2.1.

In Figure 2.1, cytological components of a colon tissue are also illustrated. The most important components for colon adenocarcinoma are glands. Glands involve relatively large luminal areas and epithelial cells surrounding these lumens. Lumens are white large regions in Figure 2.1 and they are responsible for absorption of water and minerals and secretion of mucus to waste products. Epithelial cell nuclei are stained dark purple and forms the border of the glands. Stroma is the region that connects glands and keeps them together. In stroma, there exist non-epithelial cells, which are also stained dark blue. Colon adenocarcinoma originates from epithelial cells and changes glands' structure, shape, and size. From a low grade gland to a high grade gland, the change becomes considerable. In this thesis, we focus on three classes (tissue types) in the context of diagnosis of colon adenocarcinoma. These tissue types are normal, low grade cancerous and high grade cancerous. These tissue types are exampled in Figure 2.2. As seen in Figure 2.2(a)-(b), a normal tissue does not include any cancer or there is no deformation on the structure of glands. With the beginning

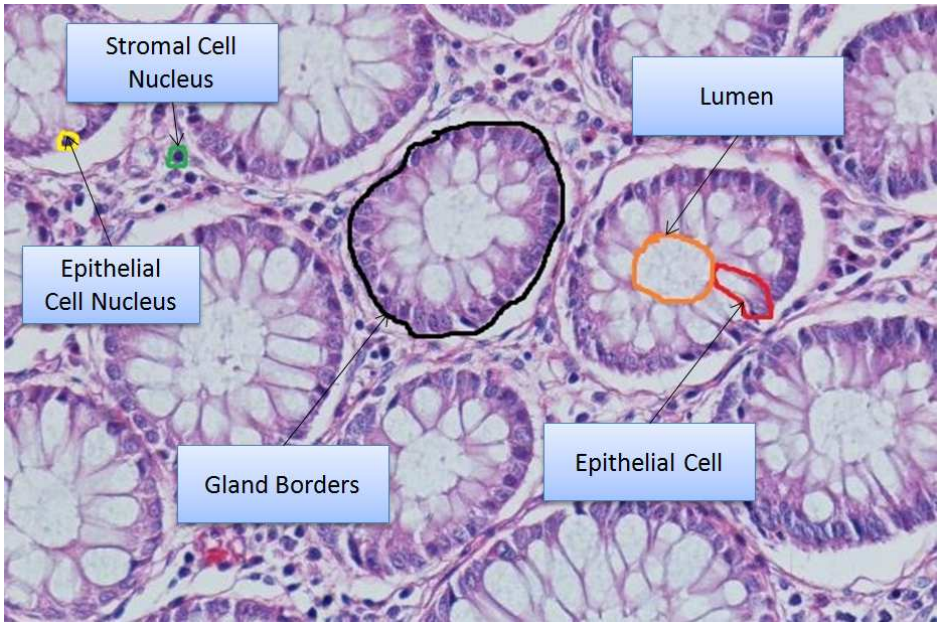
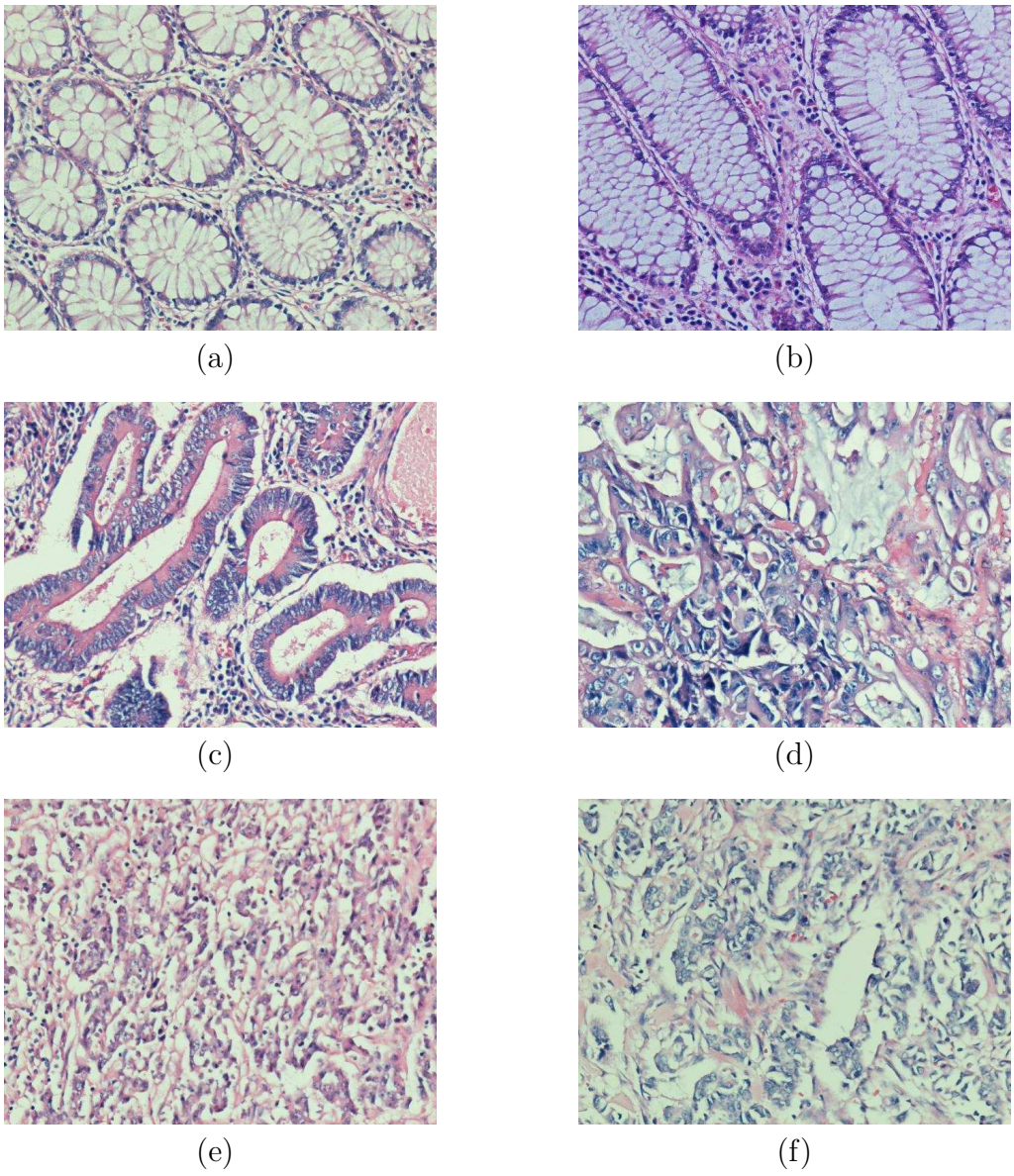


Figure 2.1: An example of a colon tissue stained with hematoxylin-and-eosin.

of the cancer, glands undergo changes in their structure; however, glands are still well to moderately differentiated in tissue images as seen Figure 2.2(c)-(d). When the cancer advances, a tissue turns into a high grade cancerous tissue. In a high grade cancerous tissue, since the deformation on the glands are too much, gland structures are only poorly differentiated, as seen in Figure 2.2(e)-(f)



(a) (b) (c) (d) (e) (f)

Figure 2.2: An illustration of example tissue images from different cancer types: (a) - (b) are examples of normal tissue, (c) - (d) are examples of low grade cancerous tissue, and (e) - (f) are examples of high grade cancerous tissue.

2.2 Automatic Cancer Diagnosis

In this section, we briefly summarize the existing computational methods proposed for automated cancer diagnosis. We group these methods into three main categories: morphological, textural, and structural.

2.2.1 Morphological Methods

Cancer causes deformation in the shapes of cells and nuclei components in a tissue. Morphological methods attempt to recognize cancer in the tissue by modeling these changes with numerical features that characterize the shape and size of these components. These features can also be aggregated to model the entire tissue by computing their average and standard deviation [60, 10]. In literature, there are various morphological features defined for quantifying cells/nuclei. Commonly used features include radius, perimeter, area, compactness, smoothness, concavity, symmetry, circularity, and eccentricity [57, 50]. Nucleocytoplasmic ratio, hyperchromasia, aninonucleosis, and nuclear deformity measures are other features for cell/nucleus quantification [60]. There are also some studies that combine the morphological features with other types of features to obtain a larger set of features. For example, in [65], the morphological features are used along with textural features for characterizing a tissue.

Extraction of morphological features requires determining the exact locations of cells beforehand, which is, however, very challenging for histopathological tissue images due to their complex nature [28]. This problem becomes a bottleneck for the morphological methods and inexact localization of the nuclei components may affect the success of these methods in negative manner.

2.2.2 Textural Methods

The existence of cancer changes texture and color of a tissue. Textural methods aim to capture these changes by extracting textural features from the tissue

image. One advantage of using textural features over using morphological and structural features is that it does not require segmentation of tissue components beforehand. There are various texture extraction methods employed for characterizing a tissue. For example, intensity and color histograms are used to model the color changes; Haralick features extracted from co-occurrence matrices utilize second order statistics of gray level intensity distribution; fractal geometry features allow quantitative representation of complex and irregular shapes; wavelet features contain information of an image at different scales. In the next subsections, we briefly mention these textural features.

2.2.2.1 Intensity and Color Histogram Features

In hematoxylin-and-eosin staining, there exist color changes when a tissue is cancerous. This can be attributed to the spread of nuclei, which are stained blue or purple, to stroma and lumina. Since, nuclei cover larger areas in cancerous tissues compared to normal ones, such tissues include large areas with low intensity values. Color histograms are used to model these changes [59]. To this end, the intensity value of each pixel in red, green, and blue channels can be discretized into N bins and the frequency of each bin is stored in a histogram.

Likewise, intensity features model the distribution of gray level intensities of pixels in tissue images. These features are extracted from gray level intensity histograms and they include mean, standard deviation, skewness, kurtosis, and entropy of the gray level distribution [69]. In order to extract these features, the histogram probability density function $h(g_i)$ is first computed such that $\sum_i h(g_i) = 1$. The intensity features are then computed on this density function. The definitions of the most commonly used intensity features are given in Table 2.1.

Table 2.1: The Definitions of the most commonly used intensity features extracted on an intensity histogram.

Mean	$m = \sum_i h(g_i) \cdot g_i$
Standard deviation	$\sigma = \sum_i (g_i - m)^2 \cdot h(g_i)$
Skewness	$s = \sum_i (g_i - m)^3 \cdot h(g_i)$
Kurtosis	$k == \sum_i (g_i - m)^4 \cdot h(g_i)$
Entropy	$e = -\sum_i h(g_i) \cdot \log_2 h(g_i)$

2.2.2.2 Co-occurrence Matrix Features

Co-occurrence matrix features are defined to characterize the spatial distribution of gray level intensity values in an image [31]. The co-occurrence matrix P is defined as a matrix that keeps the frequency of two gray levels being co-occurred in a particular spatial relationship defined by a distance d and an angle θ. Second order statistics are computed on the matrix P to obtain co-occurrence matrix features. The most commonly used co-occurrence features are listed in Table 2.2. In the table, P_{ij} stands for frequency of gray levels i and j being co-occurred, μ_x and μ_y are the means of column, row sums, and σ_x and σ_y are their standard deviations. In histopathology tissue domain, co-occurrence features are often used to characterize the texture of an entire tissue image [69, 19, 18, 36].

Table 2.2: The definitions of the most commonly used textural features extracted on a normalized co-occurrence matrix P

Energy	$\sum_{i,j} P_{ij}^2$
Entropy	$-\sum_{i,j} P_{ij} \ \log P_{ij}$
Contrast	$\sum_{i,j} P_{ij} \ (i-j)^2$
Homogenity	$\sum_{i,j} \frac{P_{ij}}{1+\lvert i-j \rvert}$
Dissimilarity	$\sum_{i,j} P_{ij} \ \lvert i-j \rvert$
Correlation	$\sum_{i,j} \frac{P_{ij} \ (i-\mu_y) \ (j-\mu_x)}{\sigma_x \cdot \sigma_y}$
Max Probability	$\max_{i,j} P_{ij}$

2.2.2.3 Gray Level Run Length Features

In statistical texture analysis, the number of combination of intensity levels determines the order of the statistics used for feature extraction. Features extracted from intensity and color histograms are examples of the first-order statistics, while co-occurrence features are examples of the second-order statistics. Gray level run length features are examples of higher order statistical texture features [1]. A gray level run is a set of consecutive pixels with the same gray level intensity in a given direction. The run length is defined as the number of pixels in a run and the run length value is the frequency of this run in an overall image. A gray level run length matrix P includes $P_{ij|\theta}$ that gives the total number of runs of a length j and a gray level intensity i at direction θ. Galloway introduces five features to be extracted from a gray level run length matrix [25] and Chu later extends them by defining two new features [11]. These features are summarized in Table 2.3, where n is the total number of pixels in the image. Let K be the total number of runs in the image such as $K = \sum_i \sum_j P_{ij|\theta}$.

Table 2.3: The definitions of gray level run length features on matrix $P_{ij|\theta}$

Short Runs Emphasis	$\sum_i \sum_j \frac{P_{ij\|\theta}}{j^2} \ / \ K$
Long Run Emphasis	$\sum_i \sum_j j^2 \ P_{ij\|\theta} \ / \ K$
Gray Level Non-uniformity	$\sum_i (\sum_j P_{ij\|\theta})^2 \ / \ K$
Run Length Non-uniformity	$\sum_j (\sum_i P_{ij\|\theta})^2 \ / \ K$
Run Percentage	$\frac{1}{n}K$
Low Gray Level Runs Emphasis	$\sum_i \sum_j \frac{P_{ij\|\theta}}{i^2} \ / \ K$
High Gray Level Runs Emphasis	$\sum_i \sum_j i^2 \ P_{ij\|\theta} \ / \ K$

In the diagnosis of cancer, gray level run length features are generally used together with other features. For instance, Bibbo *et al.* use gray level run length features along with co-occurrence and intensity features to distinguish normal and tumor nuclei. [6]. Weyn *et al.* use gray level run length features and co-occurrence features to statistically characterize nuclei of an image [66].

2.2.2.4 Fractal Geometry Features

Fractal geometry features are used at characterizing complex and irregular shaped objects in images [37]. A fractal is an object made of subobjects that are similar to the whole object in some way. Fractal dimension D gives the complexity of a fractal. Self similarity can be used to estimate the fractal dimension D. For example, let S be a self similar object which is union of N_r distinct copies of S scaled down by ratio r. We can estimate fractal dimension D of S by the expression $N_r \cdot r^D = 1$ or D can be calculated by the following equation.

$$D = -\frac{\log N_r}{\log r}$$

However, most of the natural objects do not show deterministic self similarity so fractal dimension D should be estimated. There exist different methods to estimate fractal dimension, one of the most commonly used technique is differential box-counting approach. Most of the time together with other types of features, fractal features are used to characterize texture of the histopathological images [32, 59].

2.2.2.5 Multiwavelet Features

In wavelet transform, the data is first divided into different frequency components and they are analyzed at resolution matching to their scale [29]. Multiwavelet transform uses more than one scaling function. Multiwavelet transform can preserve features such as short support, orthogonality, symmetry, and vanishing moments [56]. Zadeh *et al.* use multiwavelet features in histopathology images for automated Gleason grading of prostate tissues. To do so, using wavelet transform, they compose each tissue image into subbands. Finally, from wavelet coefficients of the subbands, they extract features such as entropy and energy to characterize the tissue image [56]. It is also possible to select distinctive subbands and use them for feature extraction [51].

2.2.2.6 Local Binary Pattern Features

Local binary patterns are a set of textural features that are used to model the texture of an image in micro-level. In this technique, a binary number is generated for each pixel in the image and these numbers are combined using a histogram. This histogram is used as a feature vector to characterize the texture of the image. To generate a binary number for a pixel, a 3×3 operator is used to compare its intensity with its neighbors' intensities; it assigns zero if a neighbor's intensity is lower than the pixel's intensity, otherwise it assigns one. These binary values for each neighbor are appended in clock-wise manner to obtain a binary number for the pixel [48]. Qureshi *et al.* extend local binary patterns by choosing neighbors of a particular pixel from those that lie on a circle which is centered on that particular pixel and has a radius of r [51]. Sertel *et al.* use local binary pattern features and co-occurrence matrix features together for detection of cancer in histopathological images [54].

2.2.3 Structural Approaches

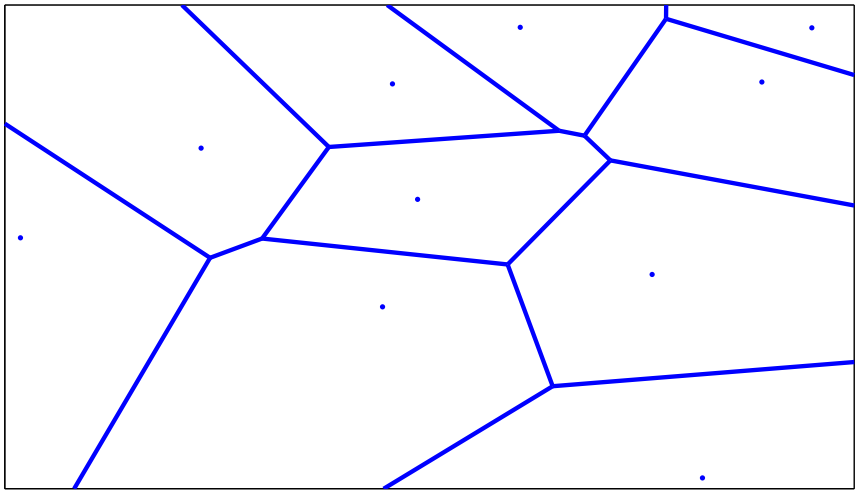
Structural approaches aim to recognize cancer from the topological changes of tissue components in cancerous tissues. Structural methods usually represent the tissue with a graph to model spatial distribution and neighborhood information of tissue components. Features extracted from these graphs are used in the automated cancer diagnosis and grading. In generation of such graphs, nuclear components are usually considered as graph nodes and edges are generated to retain spatial information among these nodes. There are various graph generation methods; these graphs include Delaunay triangulations, Voronoi diagrams, minimum spanning trees, probabilistic graphs, and weighted graphs. Recently, color graphs are used for tissue characterization. In addition to nuclear components, this approach also considers other cytological components in its graph generation. In the next subsection, we briefly mention some of these methods.

2.2.3.1 Voronoi Diagram Features

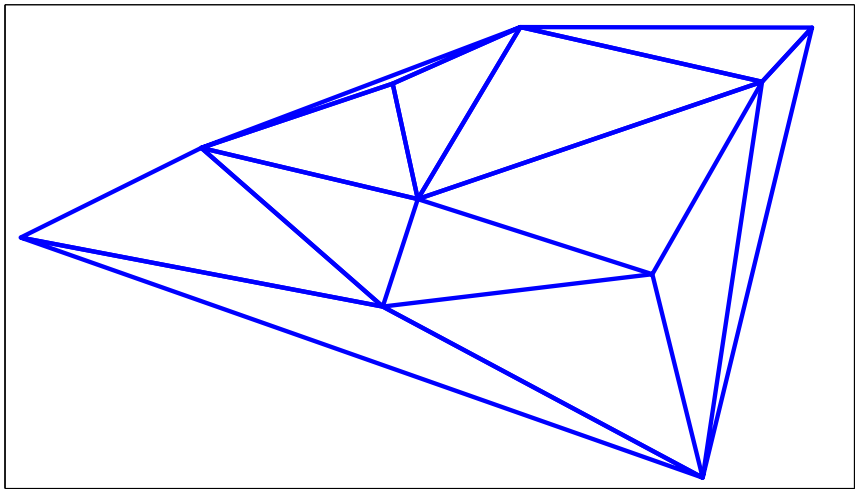
A Voronoi diagram is the partitioning of a plane into complex polygons $C = \{C_i\}_{i=1}^{N}$ with given points $O = \{o_i\}_{i=1}^{N}$ such that there exists exactly one generating point o_i for the complex polygon c_i and each point residing in the polygon c_i is closer to its generating point o_i than other generating points $O \setminus \{o_i\}$ [68]. Figure 2.3(a) shows an example of a Voronoi diagram generated on ten random points. Voronoi diagrams are also commonly used in the structural representation of histopathological images [5, 68, 58, 19]. To this end, the set of nuclear centroids on a tissue image is used as generating point set O and a Voronoi diagram is constructed on this point set. Commonly extracted features from a Voronoi diagram include the mean, standard deviation, min-max ratio, and disorder of polygon areas, the polygon perimeter lengths, and the polygon chord lengths [5].

2.2.3.2 Delaunay Triangulation Features

The Delaunay triangulation can be derived from a Voronoi diagram as they are dual of each other. The Delaunay triangulation D of a given point set $O = \{o_i\}_{i=1}^{N}$ can be constructed easily after its Voronoi diagram V is generated. For that, an edge is assigned between any two unique points, o_i and o_j, where $i \neq j$, if their corresponding polygons, c_i and c_j, in the V share a side. Figure 2.3(b) illustrates the Delaunay triangulation generated on ten random points. In histopathological image domain, many studies use Voronoi diagrams together with Delaunay triangulation to characterize the structural information of a tissue image [5, 19]. Likewise, these studies use nuclear centroids to generate Delaunay triangulation. Common features extracted from the Delaunay triangulation consist of the mean, standard deviation, min/max ratio, and disorder of the triangle edge lengths and the triangle areas.



(a)



(b)

Figure 2.3: An example of (a) a Voronoi Diagram and (b) a Delaunay Triangulation generated for ten random points.

2.2.3.3 Minimum Spanning Tree Features

A spanning tree of a connected and undirected graph G is a subgraph that connects all of the graph nodes without having any cycles. There may exist different spanning trees of the same graph. For the weighted graphs, the minimum spanning tree (MST) is defined as the one that minimizes the total spanning tree edge weights. In automatic cancer diagnosis, features extracted from minimum spanning trees are used to characterize a tissue. These features include the mean, standard deviation, min/max ratio, and disorder of the MST edge lengths [10, 17, 19, 68, 5].

2.2.3.4 Color Graphs Features

As opposed to other structural methods, which model only the spatial distribution of cell nuclei using a graph, the color graph approach is proposed to model also the distribution of other tissue components [2]. In this approach, a graph is generated considering nuclear, stromal, and luminal tissue components as graph nodes and assigning graph edges using Delaunay triangulation. In this graph representation, each edge is labeled according to the type of their end nodes. After constructing a graph, colored version of the features such as the colored average degree, colored average clustering coefficient, and colored diameter are extracted.

2.3 Limited Training Data

In supervised classification, a model is first learned from a training set and then used to classify unlabeled samples. Here the aim is to construct a good model, which is a close approximation to the true model. The difference between the constructed and true models may yield an error, which is also known as an estimation error. In order to construct a *good* model, and hence to reduce the estimation error, it is known that the model should be learned on sufficient, and usually large, amount of labeled training samples, which reflect the real data distribution [20].

On the other hand, obtaining large amount of labeled data is quite difficult, and costly, in many problem domains. Automated cancer diagnosis on histopathological images is one of such domains. To alleviate this problem, different approaches have been proposed. Active learning, which intelligently selects the samples to be labeled, is one them [40, 70] it is also possible to use semi-supervised learning, which combines labeled and unlabeled data in the construction of a classifier [39]. Another approach is to use data resampling, which is especially used to make unbalanced dataset balanced [12, 49].

2.3.1 Active Learning

Active learning methods assume that collecting unlabeled data is easy but labeling them is costly and difficult [55]. Therefore, in active learning systems, a classifier selects the samples to be labeled (and to be used in its training set) interactively. Here the aim is to obtain the highest accuracies by selecting the minimum number of samples. For that, it is possible to start with random selection of samples as an initial training set and enhance the classifier iteratively by selecting and labeling unlabeled samples. To select an unlabeled sample, Liu *et al.* calculate the distance of unlabeled samples to the SVM's hyperplane and select the sample with the highest distance. They apply this approach to gene expression data for cancer classification [40]. Yu *et al.* use a classifier to estimate the confidence score of an unlabeled sample, and decide whether or not to label the sample according to its score level. The classifier labels the sample if its confidence level is higher than a predefined threshold, the classifier leaves the sample to be labeled by an expert otherwise [70].

2.3.2 Semi-supervised Learning

Similar to active learning, semi-supervised learning assumes that a large number of unlabeled data exists in the dataset, however labeling them is difficult. This approach utilizes unlabeled samples to increase the performance of the classifier

especially where there is a limited number of labeled training samples. As opposed to active learning, semi-supervised learning does not query the label of the unlabeled samples to an expert, this is the difference between semi-supervised learning and active learning. In [39], they use ensemble of N classifiers. To refine a particular classifier, they use other classifiers to label an unlabeled sample, and they use the labeled sample in the particular classifier's training set.

2.3.3 Resampling

In resampling, samples are drawn from a sample set to obtain a new set [45]. Resampling has been used with different purposes including validation of cluster results [45] and performance evaluation of classifiers [63, 53]. Additionally, it is commonly used to alleviate negative effects of the issue of unbalanced training datasets problems [47, 22, 12, 49]. In unbalanced datasets which the number of samples is significantly different from each other among classes, classifiers tend to favor prevalent classes. To circumvent this problem, one may balance the dataset by resampling from classes with less number of samples. There are three techniques:

- Bootstrapping: In bootstrapping, each bootstrap sample is obtained by random selection from the original dataset with replacement. In this resampling technique, some samples in the original set may be selected more than once or may not be selected at all. In subsampling, random subsets $\{Y_i\}_{i=1}^{N}$ are resampled from the set X such that $Y_i \subset X$ [45].
- Jittering : If one measures the same event multiple times, these measurements may be different from each other due to the measurement error. In jittering, the main motivation is to simulate the existence of such measurement errors in selecting samples. Hence, jittering selects random samples from the original dataset and adds random noise to these selected samples [45].
- Perturbation: In bootstrapping, resampled samples are the replicates of the original samples. Likewise, in jittering, resampled samples are only

slightly different from the original ones. These resampling techniques do not simulate the differences between samples due to the intra-population variability. On the other hand, perturbation attempts to reflect estimates of intra-population variability to the original samples. In this technique, random variables are generated from a distribution that models distribution of original samples and these random variables are added to original samples to obtain perturbed samples [45]. For example, one may calculate the mean and standard deviation of the original samples, calculate random values from a normal distribution with the computed mean and standard deviation, and add these values to original samples [45].

Although these techniques increase the size of a dataset, however, for images that contain irrelevant information and a considerable amount of noise, one may want to develop techniques that do not use the entire image but some of its subregions. In this thesis, we present a novel resampling technique. This technique generates sequences that model partial regions in the tissue image and uses each of these sequences as a sample in learning and classification.

Chapter 3

Methodology

The proposed resampling-based Markovian model (RMM) relies on generating perturbed samples from each tissue image and using these perturbed samples in learning and classification. The main motivation behind the use of perturbed samples is to model variances in tissue images better even when only limited labeled data are available. The RMM includes two components: perturbed sample generation and Markov modeling. We explain these two components in the following sections.

3.1 Perturbed Sample Generation

Let I be a tissue image that is to be either classified in testing 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, $S^{(n)} = O_1^{(n)} O_2^{(n)} \dots O_T^{(n)}$. (For better readability, we will drop n from the terms unless its use is necessary. Thus, each perturbed sample is represented by $S = O_1 O_2 \dots O_T$.) These perturbed samples model partial regions of the tissue image I. There are three main steps: The first step is the selection of data points and extracting features representing them, the second step is to discretize extracted features into a set of observation symbols, and the final step is to order the observation

symbols. These steps are explained in Sections 3.1.1, 3.1.2, and 3.1.3, respectively. Figure 3.1 illustrates the general outline of the perturbed sample generation.

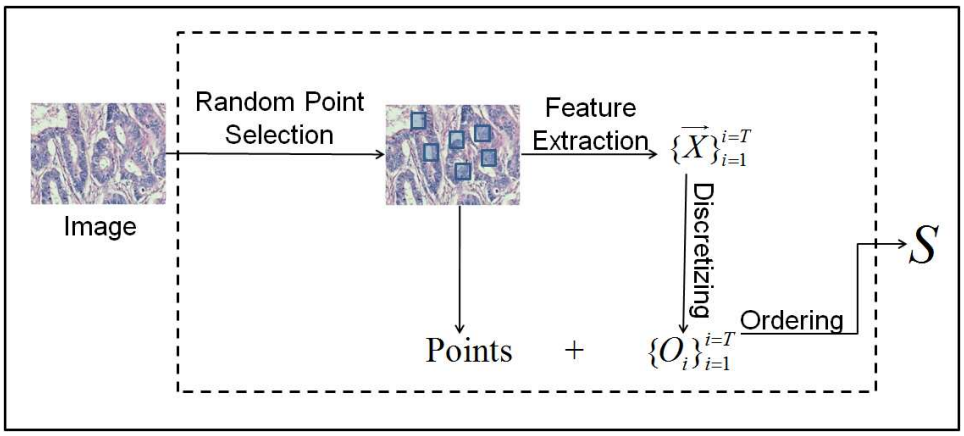


Figure 3.1: A schematic overview of the perturbed sample generation.

3.1.1 Random Point Selection and Feature Extraction

The first step of generating a sequence S from the image I is to select T data points from the image I and to characterize them by extracting features. The RMM uses random point selection to select T random points. After selection of T data points, each point is characterized by using its neighborhood pixels. To this end, we locate a window centered on each of these selected points and extract features to characterize the local image around this point. The RMM uses four features that quantify color distribution and texture of the pixels falling within this window. These four features are defined on the quantized pixels. The quantization of the pixels is mapping pixel colors to some dominant colors. In our case, tissues are stained with hematoxylin and eosin, which yields three dominant colors (white, pink and purple). However, the intensity levels of these three dominant colors show differences among tissue images due to the variability of the staining process. In quantization of pixels we map pixel colors into white, pink and purple. For that reason, the k-means algorithm is used to cluster pixel colors of the image I into three and each of these clusters are labeled as white,

pink or purple according to closeness of the center of the cluster to these three dominant colors. The first three features that the RMM uses correspond to ratios of quantized pixel colors in the window. The last feature is a texture descriptor (J-value) that quantifies how uniform the quantized pixels are distributed in space [16]. Let Z be the set of the pixels that reside in the window and Z_i be the set of the pixels that belongs to color i; in our case, since we quantize pixel colors, i might be either white, pink or purple. Let $z = (x, y) \in Z$ be a pixel. We first calculate the overall mean m of the pixels in Z, and class mean m_i for the pixels in Z_i.

$$m = \frac{\sum_{z \in Z} z}{|Z|} \tag{3.1}$$

$$m_i = \frac{\sum_{z \in Z_i} z}{|Z_i|} \tag{3.2}$$

Then, we calculate overall variation S_t and total variation of pixels belonging the same class S_w.

$$S_t = \sum_{z \in Z} ||z - m|| \tag{3.3}$$

$$S_w = \sum_{i \in \{white, pink, purple\}} \sum_{z \in Z_i} ||z - m_i|| \tag{3.4}$$

The J-value is calculated from S_t and S_w.

$$J = \frac{S_w - S_t}{S_t} \tag{3.5}$$

Note that, the RMM uses a generic feature framework and does not impose any specific feature type. One may define his/her own features and use them in the RMM. In this study, we select features that are effective and easy to compute. Besides, selected features do not introduce any external model parameters.

In this step, for the selection of data points, we use random selection by default, yet there would be other methods for the selection of these T data points. For example, one may want to identify SIFT (Scale Invariant Feature Transform) points from the image I, which are commonly used in object detection in the literature [41]. We discuss the effects of selecting of the data points randomly or via SIFT in Section 4.5.2.1.

3.1.2 Discretizing Features into Observation Symbols

After selecting data points and extracting their features, we discretize the features for the data points into K observation symbols, $\{v_k\}_{k=1}^{K}$, since we use discrete Markov processes for modeling different types of tissue formations. There are two main steps for the discretization. The first step is the unsupervised learning of the observation symbols; this step is done once and learned observation symbols are used throughout the learning and classification stages. The second step is to discretize each of the features to an observation symbol. The details of these steps are explained in the following subsections.

3.1.2.1 Learning Observation Symbols

We use k-means clustering to learn K clusters on the extracted features of the data points selected from the training images. K-means is an unsupervised clustering that partitions the data points such that the squared error between the mean of a cluster and the points in that cluster is minimized [34]. It is an iterative algorithm that minimizes the squared error in each iteration. We learn clusters by selecting 100 random data points from each training image and initialize k-means with random initial cluster centers. 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 may be selected smaller to decrease the computational time of training. In addition, note that unsupervised learning of the observation symbols provides the RMM the flexibility of automatic observation symbol learning. Hence, the RMM can easily be extended to other domains or to other feature types. One can change the feature extraction process and apply the RMM in the areas outside of histopathological image classification.

3.1.2.2 Discretizing Features

After we find $\{v_k\}_{k=1}^{K}$ observation symbols, we map extracted features of a selected point to one of the $\{v_k\}_{k=1}^{K}$ observation symbols. Let $\{m_k\}_{k=1}^{K}$ be the set of the cluster centers that each observation symbol corresponds. For each selected point P, we have a set of four features which can be denoted by X. We discretize P to the observation symbol v^* that is the label of the closest cluster center.

$$v^* = \underset{i}{\operatorname{argmin}} \; dist(X, m_i) \tag{3.6}$$

Here, we use the Euclidean distance to compute the distance between the X and each cluster center. At the end of this step, a perturbed sample is represented with a set of observation symbols, $S = \{O_i \mid O_i = v_k \text{ where } 1 \leq k \leq K \text{ for all } i\}$, but not as a sequence of them.

3.1.3 Ordering the Points

The next step is to order the data points and construct a sequence from their observation symbols. The data points are ordered as to minimize the sum of distance between the adjacent points. Formally, this ordering problem can be represented as finding $S = O_1 O_2 \dots O_T$ such that

$$\sum_{t=2}^{T} dist(P_{t-1}, P_t) \tag{3.7}$$

is minimized. Here $dist(u, v)$ represents the Euclidean distance between the points u and v and O_t is the observation symbol defined for the point P_t. This formulated problem indeed corresponds to finding the shortest Hamiltonian path among the given points, which is known as NP-complete. Thus, the proposed method uses a greedy solution for ordering. This greedy 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 $dist(P_{t-1}, P_t)$. In Figure 3.2, we illustrate a quantized image and a perturbed sample generated from this quantized image. Note that for simplicity, we select length of the sequence as 40 and number of the distinct states as 10. We repeat this process to obtain N sequences.

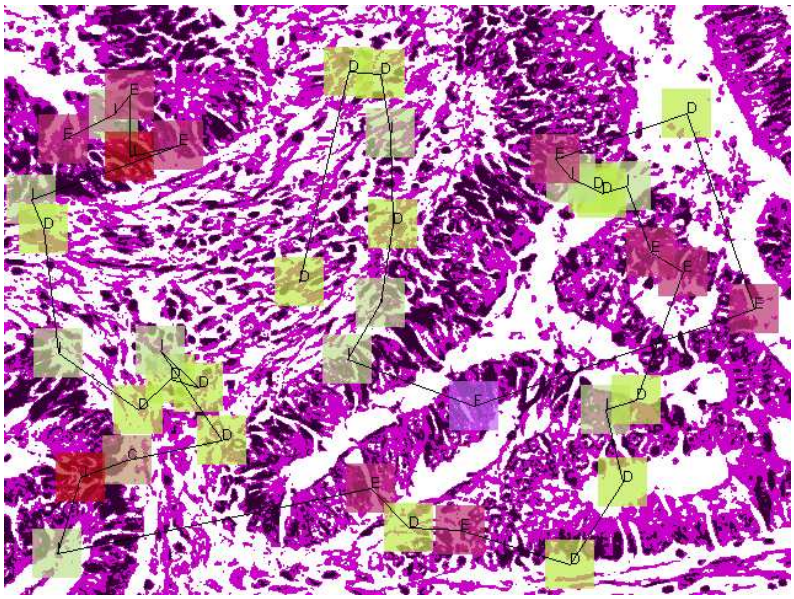
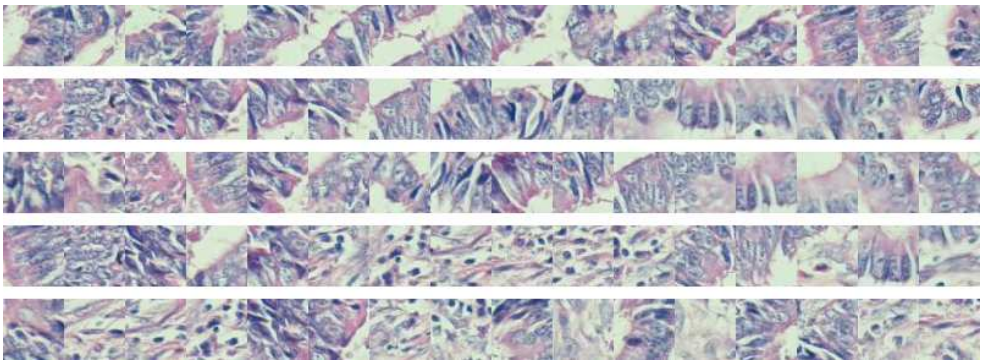


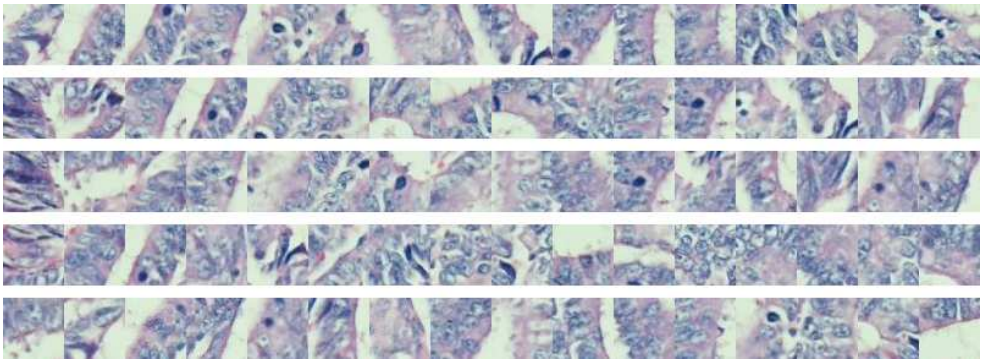
Figure 3.2: A quantized image and a sample sequence.

At the end of this step, we obtain N sequences, each representing a perturbed image. These sequences are expected to model variances in tissue images better. To illustrate the reason behind this, let us consider the tissue images in Figures 1.2(c) and 1.2(d). Although they show variances at the pixel level due to their irrelevant regions, these images are indeed very similar to each other in terms of their biological context. Figures 3.3(a) and 3.3(b) show some sequences generated from these images; here a data point is represented with its window, in which its features are extracted. As observed in these figures, it is possible to obtain similar sequences for these two images; the first three sequences of Figure 3.3(a) are visually similar to those of Figure 3.3(b). In our proposed RMM, we anticipate to have some of such similar sequences provided that a large number of sequences are generated.

Generating perturbed samples increases number of samples in the dataset. Besides, it also increases the diversity. For instance, if we extract features using an entire image, we would obtain one feature vector representing the entire image. However with perturbed sample generation, we model different parts of the image



(a)



(b)

Figure 3.3: Sequences generated for the tissue images given in Figures 1.2(c) and 1.2(d). Similar sequences could be obtained for these two images even though they show variances at the pixel level due to their irrelevant regions.

and use the information from these different parts for learning and classification. Moreover, although each of these sequences is generated from the same image, they are different from each other. For instance in Figure 3.3(a) or 3.3(b), the perturbed samples are generated from the same image and they are directly related to the same image, however as you observe, they are different from each other and some of these sequences even do not resemble to each other. Hence, by using perturbed samples, we increase the diversity in the training set.

3.2 Markov Modeling

A Markov process models the state of a system with a random variable X which changes with time. For instance, X_1 gives the state of the system at time one. A

Markov model is a stochastic process, where the value of the X_t depends on its previous s states. For example, in the first-order Markov model, s becomes one; i.e., the future behavior of the system (state of system) depends on its previous state. A discrete Markov model is a Markov process whose state space is a finite set. There are two common types of Markov models: observable and hidden Markov models. In an observable Markov model, the states are equivalent to the observations; that is with the knowledge of observation, we can precisely infer the state. In a hidden Markov model, the observations are related to underlying state, but the knowledge of observation is insufficient to precisely infer the underlying state.

In the proposed RMM, we model the perturbed samples each of which is a sequence of observation symbols $S = O_1 O_2 \dots O_T$ with Markov models. In addition, we assume one-to-one correspondence with observation symbols and states, and a finite set of the observation symbols. These assumptions allow us to use a m'th order observable discrete Markov model. In this study, we use $m = 1$, so observation symbol O_t is only dependent to its previous observation symbol O_{t-1}.

$$\begin{aligned} &P(O_t = v_i \quad | \quad O_{t-1} = v_j,\ O_{t-2} = v_k, \dots) = \\ &P(O_t = v_i \quad | \quad O_{t-1} = v_j) \end{aligned} \tag{3.8}$$

In this study, we use Markov modeling since it is one of the simplest and most effective ways of modeling sequences. There are also other possible methods for sequence modeling such as hidden Markov models and recurrent neural networks. We believe that these methods can work well in the proposed method, provided that their parameters are correctly estimated on the training data. However, since Markov modeling provides a fairly accurate tool for our purpose and because of its simplicity, we prefer using it.

3.2.1 Learning the Parameters of a First Order Observable Discrete Markov Model

For each class C_m, we train a different Markov model. Each 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

$$\pi(v_i \mid C_m) = P(O_1 = v_i \mid S \in C_m) \tag{3.9}$$

$$a(v_i, v_j \mid C_m) = P(O_{t+1} = v_j \mid O_t = v_i \text{ and } S \in C_m) \tag{3.10}$$

The number of states K_m is the same for every Markov model and equal to the number of observation symbols. For learning the probabilities Π_m and A_m, a new training set, $D_m = \{S^{(u)} \mid S^{(u)} \in C_m\}$, is formed generating N perturbed samples from each training image that belongs to the class C_m. Using this new training set, the probabilities are learned by maximum likelihood estimation. The process is illustrated in Figure 3.4.

$$\pi(v_i \mid C_m) = \frac{\#\{S^{(u)} \text{ such that } O_1^{(u)} = v_i\}}{\#\{S^{(u)}\}} \tag{3.11}$$

$$a(v_i, v_j \mid C_m) = \frac{\sum_{t=1}^{T-1} \#\{S^{(u)} \text{ such that } O_t^{(u)} = v_i, O_{t+1}^{(u)} = v_j\}}{\sum_{t=1}^{T-1} \#\{S^{(u)} \text{ such that } O_t^{(u)} = v_i\}} \tag{3.12}$$

In Equations 3.11 and 3.12, $\#\{.\}$ denotes "'number of"' function. In these equations, if there is no occurrence of a particular event in the training data, then the formulas yield zero probability for those events. For example, if there is no subsequent observation of v_i and v_j in any sequence in the training set, the state transition probability $a(v_i, v_j)$ will be zero. This case may not be desired especially when there is limited data for training, since many probabilities will be zero, although the occurrence of these events would be plausible. Zero probability

is indeed very ambitious, as not having an observed occurrence of an event does not really mean it will not happen. In order to handle this problem, in this study, we use additive smoothing [8] with $\alpha = 1$. Additive smoothing assumes an occurrence of each event from the outset. The α parameter gives initial frequency to each possible event. Therefore, the initial and transition probabilities can be computed as

$$\pi(v_i \mid C_m) = \frac{\#\{S^{(u)} \text{ such that } O_1^{(u)} = v_i\} + \alpha}{\#\{S^{(u)}\} + \alpha \cdot K} \tag{3.13}$$

$$a(v_i, v_j \mid C_m) = \frac{\sum_{t=1}^{T-1} \#\{S^{(u)} \text{ such that } O_t^{(u)} = v_i, O_{t+1}^{(u)} = v_j\} + \alpha}{\sum_{t=1}^{T-1} \#\{S^{(u)} \text{ such that } O_t^{(u)} = v_i\} + \alpha \cdot K} \tag{3.14}$$

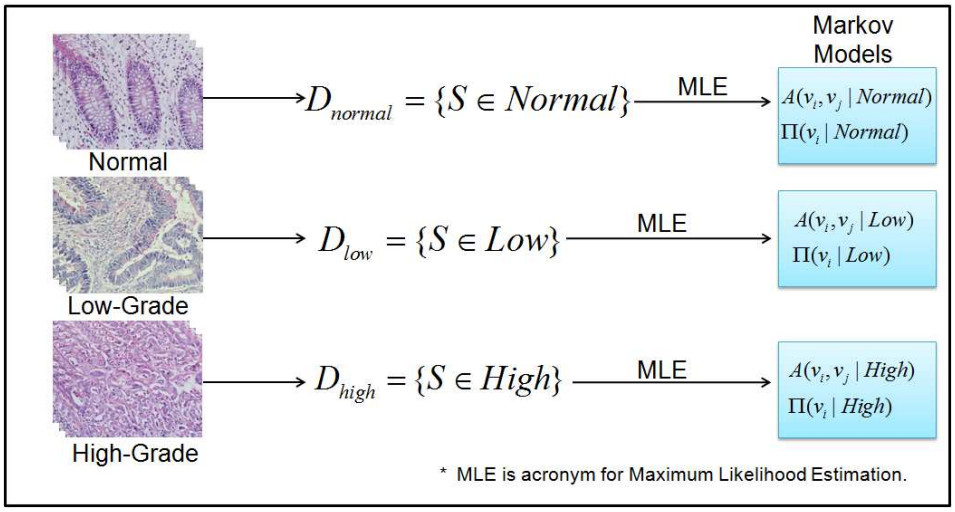


Figure 3.4: A schematic overview of the proposed resampling-based Markovian model (RMM) for learning the model parameters.

3.2.2 Classification

The classification of a given image I is done using its perturbed samples. For each perturbed sample $S \in \{S_i\}_{i=1}^{N}$ of I, the posterior probability of each class C_m is

computed and the class that maximizes these posterior probabilities is selected. Posterior probability $P(C_m \mid S_i)$ is the probability of the perturbed sample S_i belonging to class C_m. Subsequently, the class C^* of the image I is found using a majority voting scheme that combines the selected classes of the perturbed samples of I.

$$\delta_{ki} = \begin{cases} 1 & \text{if } k = \text{argmax}_m \, P(C_m \mid S_i) \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

$$C_* = \underset{j}{\text{argmax}} \sum_{i=1}^{N} \delta_{ji} \tag{3.16}$$

The posteriors $P(C_m \mid S)$ are calculated by the Bayes rule:

$$P(C_m \mid S) = \frac{P(S \mid C_m) \cdot P(C_m)}{P(S)} \tag{3.17}$$

where

$$P(S) = \sum_{m} P(S|C_m) \cdot P(C_m) \tag{3.18}$$

In the Bayes rule, there are two unknowns, the first one is class probabilities $P(C_m)$. In this study, we assume that each class is equally likely that is $P(C_1) = P(C_2) = ... = P(C_i) = ... = P(C_m)$. The other unknown is the class likelihood $P(S \mid C_m)$. Once Π_m and A_m are learned on the training sequences, class likelihood $P(S \mid C_m)$ can be written as

$$P(S \mid C_m) = \pi(O_1 \mid C_m) \prod_{t=1}^{T-1} a(O_t, O_{t+1} \mid C_m). \tag{3.19}$$

Since we assume the equal class prior probabilities, the class m that maximizes class likelihood $P(S \mid C_m)$ will also maximize posterior probability $P(C_m \mid S)$. The steps of the RMM to classify an unseen image are given in Figure 3.5.

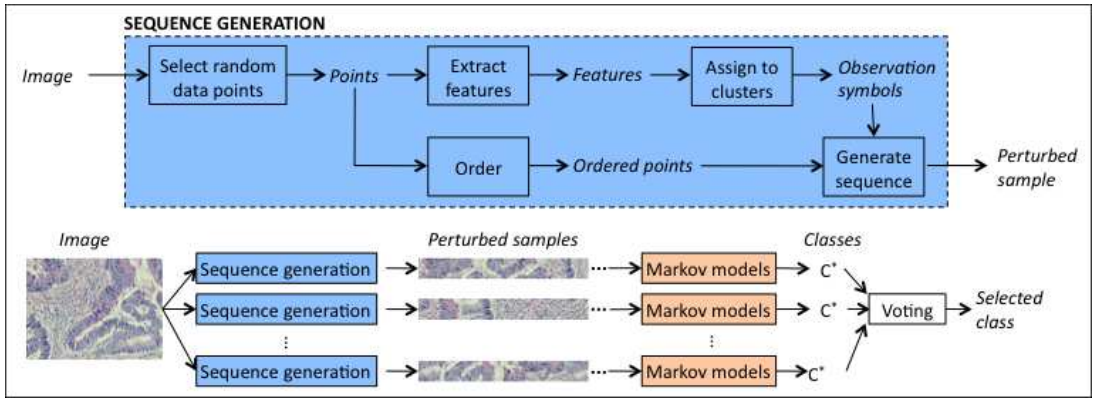


Figure 3.5: A schematic overview of the proposed resampling-based Markovian model (RMM) for classifying a given image.

Chapter 4

Experiment Results

In this chapter, we first give the details of the dataset and explain the methods that we compare with the proposed RMM. Then we describe the cross-validation technique that is used in parameter selection of the RMM and other methods. Next, we present the results of our experiments conducted with different training data sizes to understand how accurate and stable the RMM is against other algorithms. Finally, we discuss the results and give the detailed parameter analysis of the RMM.

4.1 Dataset

The dataset used in the experiments contains 3236 microscopic images of colon tissues of 258 randomly selected patients from the Pathology Department archives in Hacettepe University School of Medicine. The tissues are stained with hematoxylin and eosin and their images are taken with a Nikon Coolscope Digital Microscope using 20× microscope objective lens at 480×640 image resolution. This magnification level is high enough to ensure that different regions in a tissue image are in the same class in the context of cancer diagnosis. In addition, it is low enough to demonstrate many instances of glands in the same image.

We randomly divide the patients into two 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[1]. 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. As you may notice, the dataset we use is unbalanced and may favor low-grade cancerous class. In order to obtain unbiased results, we resample normal and high-grade cancerous tissue images randomly until we have balanced number of samples for all class types.

4.2 Comparisons

To investigate the effectiveness of the 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 perturbed sample 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. In these algorithms, we use SVM with a linear kernel for classification.

4.2.1 Algorithms with Similar Features

4.2.1.1 GridBasedApproach

First, we implement a grid-based counterpart of our method. In this approach, an image is divided into grids, the same RMM features are extracted for the grids,

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

and the grid features are averaged all over the tissue image. The details of the extracted features are given in Section 3.1.1. Then, a support vector machine (SVM) with a linear kernel is used for learning and classification tasks. 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.

4.2.1.2 VotingApproach

In this approach, we modify the previous grid-based approach so that it includes resampling-based voting. This approach 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 on the tissue image.

4.2.1.3 BagOfWordsApproach

The previous two approaches directly use the extracted grid features, without discretizing the grids. The *BagOfWordsApproach* discretizes the grids into K observation symbols 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 visual word, and uses the frequency of these visual words to characterize the image. This classifier treats an image as a collection of regions and ignores spatial information of these regions.

4.2.2 Algorithms with Different Features

4.2.2.1 IntensityHistogramFeatures

We calculate first-order histogram features over a tissue image. The *IntensityHistogramFeatures* include mean, standard deviation, kurtosis, and skewness values calculated on the intensity histogram of a gray-level tissue image [69]. To reduce the effects of noise or small intensity differences, pixel intensities are quantized into N bins.

4.2.2.2 IntensityHistogramFeaturesGrid

These features are the same with the previously defined intensity histogram features. The difference is that instead of extracting a single histogram for an entire image, the image is divided into grids and first-order histogram features are computed for every grid. Then average of these features is used to characterize the entire tissue.

4.2.2.3 CooccurrenceMatrixFeatures

We compute the *CooccurrenceMatrixFeatures* that are second-order statistics of gray level intensities in the tissue image. These features are energy, entropy, contrast, homogeneity, correlation, dissimilarity, inverse difference moment, and maximum probability derived from a gray-level co-occurrence matrix of an entire image [21, 31]. In our experiments, we define co-occurrence matrices for eight different directions and take their average to obtain rotational invariant co-occurrence matrix M_{avg}, and calculate the features on M_{avg}. Likewise, gray-level pixel intensities are quantized into N bins to lessen the effects of noise and small intensity differences.

4.2.2.4 CooccurrenceMatrixGrid

Likewise, we calculate the *CooccurrenceMatrixFeaturesGrid* features that are the grid based version of the *CooccurrenceMatrixFeatures*. In *CooccurrenceMatrixFeaturesGrid*, we divide the image into grids and the aforementioned co-occurrence matrix features are calculated for each grid. The average of these features is taken to represent an entire image.

4.2.2.5 ColorGraphFeatures

They are structural features extracted on color graphs [2]. In a color graph, nodes correspond to tissue components (nuclear, stromal, and luminal components) that are approximately located by an iterative circle-fit algorithm [61] and edges are defined by a Delaunay triangulation constructed on 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 structural features. Note that the circle-fit algorithm uses two parameters for locating the nodes.

4.2.2.6 DelaunayTriangulationFeatures

Another type of structural features we calculate is the Delaunay triangulation features. This set of features 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 [17].

4.3 Parameter Selection

The proposed resampling-based Markovian model (RMM) has four external parameters:

- *WinSize* is the size of the window in which the features of a sampled point are defined,
- *StateNo* is the number of states in a Markov model,
- *SeqLen* is the length of an observation sequence,
- *SeqNo* is the number of sequences (perturbed samples) generated for each image.

Note that the number of states and observation symbols is the same in observable Markov models. We tune these parameters with three fold cross-validation on the training set. To do that, we first determine a set of plausible values for each parameter. Then, we create a list that consists of all possible combinations of the aforementioned values for each parameter. Afterwards, we measure the success of each parameter combination in the list with three fold cross-validation on the training set and select the parameter combination that has the highest accuracy. In three fold cross-validation, to measure the success of given parameters, we first divide the training set into three equal size subsets, and each time we train the classifier with two of the subsets and given parameters, then test the classifier with the other subset. Since, we use three fold cross-validation, we repeat this process three times; for each time we test a different subset. Then we get three accuracy values and take the average of these values to obtain the overall success of the given parameters.

In our experiments, we consider all possible combinations of the following parameter sets: $winSize = \{10, 20, 40, 80\}$, $stateNo = \{4, 8, 16, 32, 64\}$, $seqLen = \{10, 25, 50, 100, 150\}$, and $seqNo = \{10, 25, 50, 100, 150\}$. Using three fold cross-validation on training images, we select the parameters as $winSize = 40$, $stateNo = 64$, $seqLen = 100$, and $seqNo = 100$. The other algorithms

Table 4.1: The parameters of the algorithms together with their values considered in cross validation.

GridBasedApproach	Grid size = $\{10, 20, \mathbf{40}, 80\}$ C = **70**
VotingApproach	Grid size = $\{10, 20, \mathbf{40}, 80\}$ Number of grids = $\{10, 25, 50, \mathbf{100}\}$ Trial number = $\{\mathbf{10}, 25, 50, 100\}$ C = **350**
BagOfWordsApproach	Number of words = $\{4, 8, 16, 32, \mathbf{64}\}$ Grid size = $\{\mathbf{10}, 20, 40, 80\}$ C = **1**
IntensityHistograms	Bin number = $\{4, 8, \mathbf{16}, 32\}$ C = **200**
IntensityHistogramGrids	Bin number = $\{4, \mathbf{8}, 16, 32\}$ Grid size = $\{\mathbf{10}, 20, 40, 80\}$ C = **550**
CooccurrenceMatrices	Bin number = $\{4, \mathbf{8}, 16, 32\}$ Distance = $\{\mathbf{5}, 10, 20, 40\}$ C = **900**
CooccurrenceMatrixGrids	Bin number = $\{4, \mathbf{8}, 16, 32\}$ Distance = $\{5, \mathbf{10}, 20, 40\}$ Grid size = $\{10, 20, \mathbf{40}, 80\}$ C = **30**
ColorGraphs	Structuring element size = $\{\mathbf{3}, 5, 7, 9\}$ Circle area threshold = $\{5, \mathbf{10}, \ldots, 50\}$ C = **3**
DelaunayTriangulations	Structuring element size = $\{\mathbf{3}, 5, 7, 9\}$ Circle area threshold = $\{5, \mathbf{10}, \ldots, 50\}$ C = **900**

have also parameters, which are listed in Table 4.1. In addition to these, they have the SVM parameter C as they use SVM classifiers with linear kernels [7]. Similarly, we use three fold cross-validation on training images to select the parameters of each algorithm. The candidate values of each parameter are given in Table 4.1. For all algorithms, the same set is considered for the SVM parameter: $C = \{1, 2, \ldots, 9, 10, 20, \ldots, 90, 100, 150, \ldots, 950, 1000\}$. The selected parameters for these algorithms are highlighted in bold in Table 4.1.

4.4 Test Results

As tissue images typically contain a considerable amount of variance, tissue classification systems 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. Note that, 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. In order to measure the success of the RMM in limited training data, we conduct our experiments using all available training data as well as using less training data. For that purpose, we randomly divide the training dataset into smaller subsets such that each subset includes P percent of the data in the original set. For all algorithms, we repeat the experiments when P is selected as 2.5, 5, 10, 25, and 50 percents. Since there are more than one subset for a selected P value (e.g., there are 20 different subsets when $P = 5$ percent), we consider all these 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 different runs in average computation.

In Tables 4.2, 4.3, and 4.4, we report the average test set accuracies obtained by the algorithms when we use all available training samples ($P = 100$ percent) and a partial set of available training samples ($P = 10$ and $P = 5$ percent) respectively. The results show that the RMM improves the accuracy of the other algorithms; the McNemar's test gives that the overall accuracy improvement is statistically significant with $\alpha = 0.05$ for all 40 runs. It is also observed that the algorithms that use grid-based aggregation to model a tissue image usually performs 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 Figure 1.2). The RMM, which can also be considered as an aggregation method, further improves these grid-based

algorithms.

Table 4.2: Classification accuracies on the test set and their standard deviations. The results are obtained when all training data are used (when $P = 100$ percent).

		Normal	Low	High	Overall
Similar Features	RMM	95.64 (±0.18)	87.77 (±0.32)	88.56 (±0.39)	90.32 (±0.18)
	GridBasedApproach	91.65	85.31	85.60	87.31
	VotingApproach	90.02	85.43	85.99	86.93
	BagOfWordsApproach	94.91	87.32	76.65	87.94
Different Features	IntensityHistograms	80.65	69.55	70.04	73.05
	IntensityHistogramGrids	78.82	74.17	78.60	76.32
	CooccurrenceMatrices	83.10	81.64	77.82	81.47
	CooccurrenceMatrixGrids	87.58	84.12	85.60	85.43
	ColorGraphs	92.67	82.46	86.38	86.24
	DelaunayTriangulations	89.61	71.56	87.55	79.71

The RMM yields better accuracies than the *GridBasedApproach* and the *VotingApproach*, which do not make use of the discretized grids in their classification. This indicates the usefulness of state definition of the RMM. In addition, the *GridBasedApproach* and the *VotingApproach* aggregate heteregenous regions by averaging the features extracted from grids. However, the RMM does not aggregate them, but uses them separately in a sequence. This would be another advantage of the RMM over the *GridBasedApproach* and the *VotingApproach*. Besides, comparing the RMM against the *VotingApproach*, the results show that generating perturbed sample 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*. This shows the effectiveness of using resampling-based voting

Table 4.3: Classification accuracies on the test set and their standard deviations. The results are obtained when limited training data are used (when $P = 10$ percent).

		Normal	Low	High	Overall
Similar Features	RMM	95.22 (±0.58)	89.45 (±1.99)	86.46 (±2.94)	90.75 (±0.66)
	GridBasedApproach	90.31 (±3.00)	84.30 (±3.28)	82.96 (±3.13)	85.94 (±0.77)
	VotingApproach	89.47 (±3.03)	84.67 (±3.10)	81.75 (±3.62)	85.68 (±0.88)
	BagOfWordsApproach	93.73 (±1.88)	90.20 (±2.04)	61.40 (±9.18)	86.64 (±1.42)
Different Features	IntensityHistograms	79.04 (±3.58)	69.64 (±5.97)	63.23 (±8.84)	71.51 (±3.00)
	IntensityHistogramGrids	77.70 (±3.33)	73.35 (±5.61)	75.25 (±3.80)	75.00 (±2.73)
	CooccurrenceMatrices	78.88 (±4.09)	81.15 (±4.02)	70.23 (±8.32)	78.69 (±1.61)
	CooccurrenceMatrixGrids	83.77 (±3.08)	83.95 (±3.05)	81.87 (±4.89)	83.56 (±1.84)
	ColorGraphs	88.37 (±3.35)	84.23 (±2.55)	75.64 (±5.37)	84.12 (±1.68)
	DelaunayTriangulations	86.80 (±1.91)	72.75 (±7.63)	74.94 (±8.62)	77.44 (±3.12)

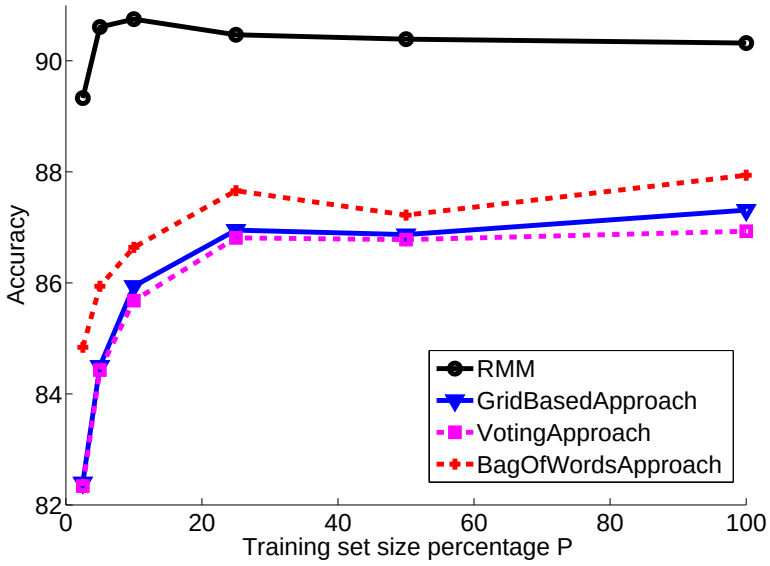
and utilizing neighbourhood information of regions, since the *BagOfWordsApproach* does not apply resampling-based voting and does not retain neighborhood information of the regions. This improvement is especially observed for correct classification of high-grade cancerous tissues; as a future research aspect of this work, one could work on incorporating the proposed framework into a bag-of-words approach. Additionally, as opposed to the RMM, none of the algorithms represent an image using perturbed image sequences. Hence, these results also indicate the importance of the sequence representation in the RMM.

Other algorithms that use different textural and structural features perform worser than the RMM. The *CooccurrenceMatrixGrids* and *ColorGraphs* are the most successful ones among them. However, as the results show, their accuracy is less that of the RMM. This indicates the success of the RMM over other methods.

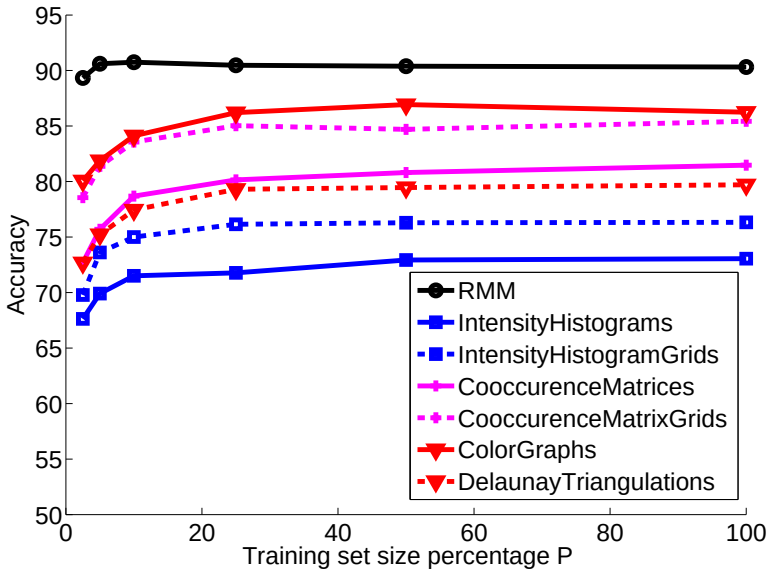
Table 4.4: Classification accuracies on the test set and their standard deviations. The results are obtained when limited training data are used ($P = 5$ percent).

		Normal	Low	High	Overall
Similar Features	RMM	94.69 (±1.37)	90.76 (±2.84)	82.32 (±5.27)	90.61 (±0.88)
	GridBasedApproach	87.25 (±5.39)	85.12 (±4.97)	77.20 (±8.15)	84.50 (±2.56)
	VotingApproach	86.75 (±5.60)	85.67 (±4.90)	75.91 (±8.19)	84.43 (±2.46)
	BagOfWordsApproach	92.31 (±2.84)	90.59 (±4.05)	58.48 (±9.25)	85.94 (±2.05)
Different Features	IntensityHistograms	77.08 (±4.84)	69.54 (±9.44)	57.43 (±10.19)	69.91 (±4.68)
	IntensityHistogramGrids	75.67 (±6.21)	73.35 (±7.65)	70.56 (±6.43)	73.61 (±4.10)
	CooccurrenceMatrices	75.17 (±6.75)	79.26 (±6.37)	65.04 (±11.12)	75.70 (±3.75)
	CooccurrenceMatrixGrids	81.45 (±5.01)	82.64 (±5.39)	77.02 (±9.44)	81.37 (±2.96)
	ColorGraphs	85.22 (±4.38)	85.79 (±5.26)	62.70 (±8.62)	81.89 (±3.23)
	DelaunayTriangulations	82.03 (±6.20)	75.31 (±8.68)	61.93 (±9.13)	75.22 (±4.43)

If we compare the co-occurrence matrix and intensiy features, we observe that the use of the co-occurrence matrix features are more effective than the intensity features. This can be attributed to potential of texture features capturing more information than the intensity features. The *ColorGraphs* features represent topological structure of tissue components. It requires correct localization of tissue components to model the topological structure, however this is a hard task in histopathological image analysis domain.

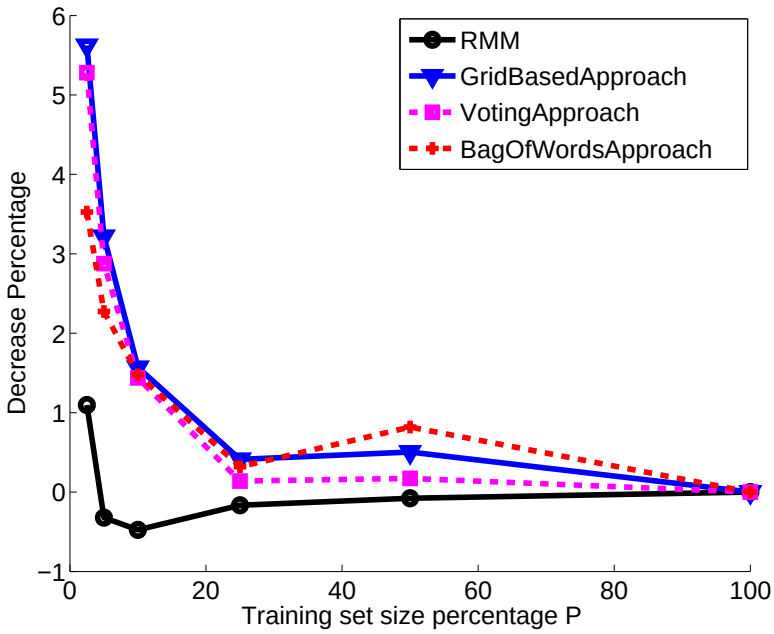


(a)

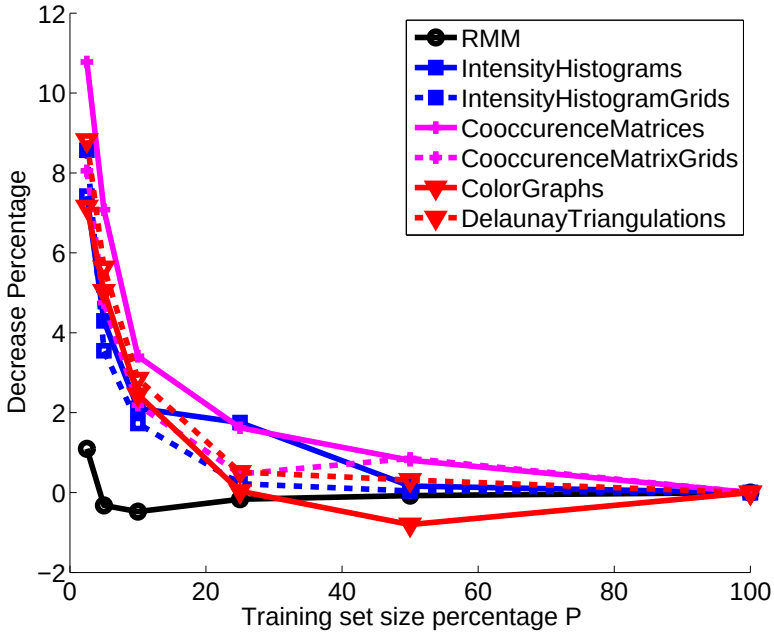


(b)

Figure 4.1: 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.



(a)



(b)

Figure 4.2: Performance decrease 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.

Figure 4.1 plots the test set accuracies and Figure 4.2 plots percentage of performance decrease for the classifiers when the partial training set is used for learning (i.e., when $P = \{2.5, 5, 10, 25, 50\}$ percents). These results are also explicitly given in Tables 4.2, 4.3, and 4.4 for $P = 100$, $P = 10$, and $P = 5$ percent of the entire training set is used. These results show that the test set accuracies decrease with the decrease in the number of training samples. For the other methods, this decrease becomes noticeable when $P \leq 25$ percent (i.e., when ≤ 411 samples are used for training). However, the proposed RMM is able to keep the test accuracy high even when 5 percent 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 higher, this slightly increases the overall accuracy.

The high performance of the RMM is attributed to the following property of the algorithms. The other algorithms used in the experiments do not attempt to vary training images for better generalizations. They just use the available training images (their features) in their current form for learning. However, the proposed 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 (samples) it generates from an individual training image. In the experiments, we make use of this property of the RMM, adjusting 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 $P = 5$ percent). This property becomes especially important when the number of training images becomes smaller and smaller. This may be one of the major reasons behind obtaining stable accuracy results until $P = 5$ percent. When it becomes 2.5 percent, a decrease is observed also for the RMM. This is due to a relatively higher accuracy decrease in high grade cancerous tissues. The number of high grade cancerous tissue images is relatively smaller in the training set (6.9 images on the average when $P = 2.5$ percent) and resampling is not able to sufficiently vary the data with such a small size of these images.

4.4.1 Unbalanced Data Issue

In unbalanced datasets, where samples of one or more classes outnumbers members of the other classes, a classifier would obtain a high accuracy by classifying every sample as one of the prevalent classes. In our case, classifiers tend to ignore high-grade cancerous tissues since there are less number of high-graded cancerous tissue with respect to other classes. A good classifier should perform efficiently in all tissue types. ROC curves are useful for domains with unbalanced datasets and unequal classification errors. We also give ROC curves for the RMM in Figure 4.4.1. ROC curves are calculated on two class domains. For that, we consider three two-class classifications problems, each of which distinguishes the images of one class from those of the others, and obtain a ROC curve for each of these problems. For example, we consider normal-vs-not classification, in which normal tissue images belong to the normal class and the other tissue images (low-grade cancerous and high-grade cancerous tissue images) belong to the not class. For this classification, we obtain a ROC curve as follows: First, for each image, we compute the ratio of its sequences that are labeled as normal. Then, if this ratio is greater than a threshold, we label the image as normal; otherwise, we label it as not. Using different thresholds from 0.0 to 1.0, we obtain a ROC curve for normal-vs-not classification. Similarly, we obtain ROC curves for lowGrade-vs-not and highGrade-vs-not classifications. We present these ROC curves in the same figure below. Please note that these ROC curves are obtained by averaging the results of all of our 40 runs and when all training data are used (i.e., P = 100 percent). The area under curve (AUC) indicates performance of the classifier [23]. Therefore, we may say these curves show that the RMM is successful in classification of all tissue types.

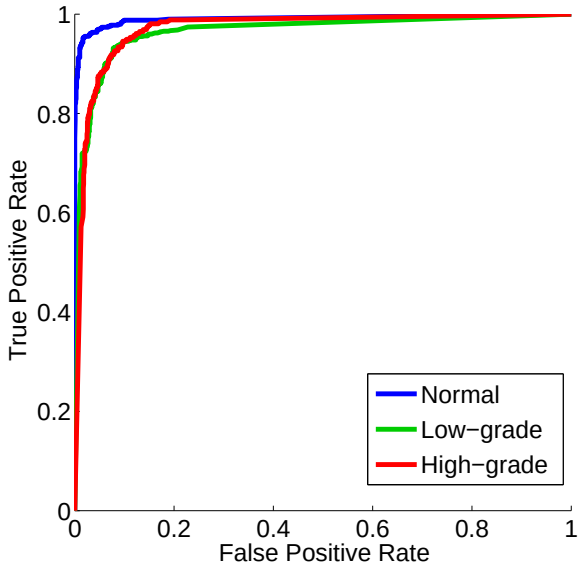


Figure 4.3: Three ROC curves of RMM which are for normal, low level cancerous and high level cancerous classes.

4.5 Analysis

The RMM uses several explicit parameters and implicit choices. In this section, we analyze the effect of these parameters to the success of the algorithm. As mentioned before, explicit parameters are selected with three fold cross-validation. These parameters are the window size, the number of states, the sequence length, and the number of sequences. In addition to these explicit parameters, the RMM includes other implicit choices. Implicit choices are taken intuitively, as a system design choice. These implicit choices are:

- *Point Selection Algorithm*: It is the choice of the algorithm that selects points on image. We use random selection by default.
- *Selected Features*: It is the choice of the features that will characterize the texture of the pixels in the window that are located around selected points.

- *Ordering Algorithm*: It is the choice of the algorithm that orders randomly selected points to generate sequences.
- *Order of a Markov Model*: It is the order of a discrete Markov model.

4.5.1 Explicit Parameters

The RMM has four external parameters: the window size, the number of states, the sequence length, and the 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 Figure 4.4 for the window size and the number of states and in Figure 4.5 for the sequence length and the number of sequences.

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 Figure 4.4(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 perturbed 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

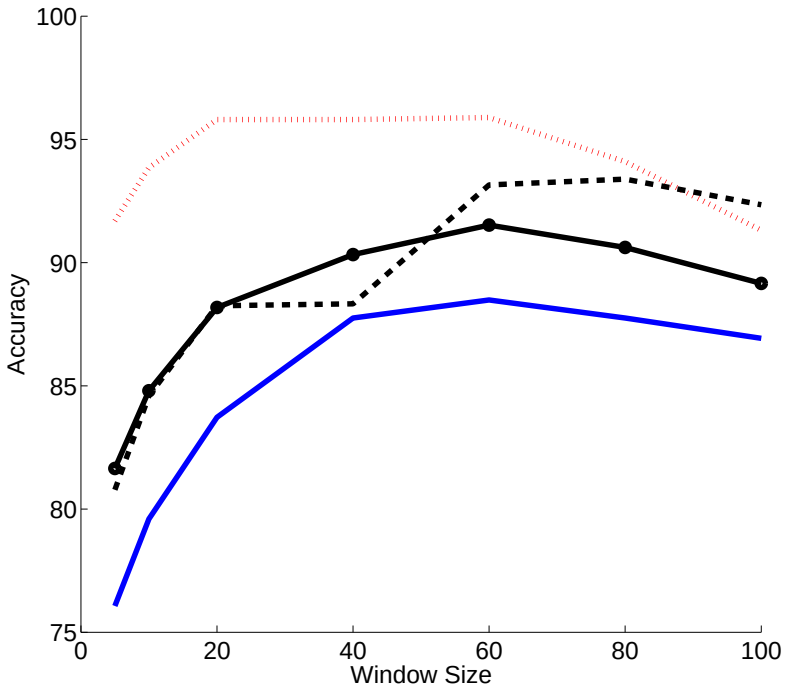
controls the number of perturbed samples generated to represent a tissue image. If it is selected too small, there is a risk of not obtaining representative samples from the image (this is closely related to our interpretation illustrated in Figure 4.5). Additionally, the number should be more than one to employ the voting scheme in classification. Although, it is not seen from the experiments, we also observe that sequence length and the number of sequences are compensating each other. This is reasonable since one long sequence will yield almost the same effect of many short sequences in learning. This is the reason of stability seen in Figure 4.5(a) and (b) with respect to changing values of these parameters.

4.5.2 Implicit Choices

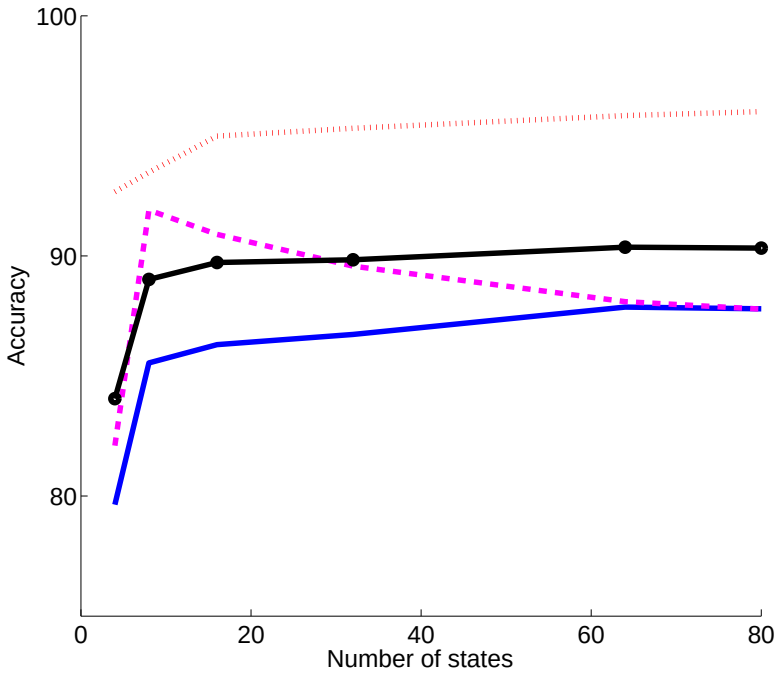
Implicit choices of the RMM are the point selection algorithm, selected features to represent texture of the grids, ordering algorithm to order states, and the order of Markov models. In the analysis of these implicit choices, we examine each implicit choice separately by observing the change in the performance of the RMM with respect to different choices.

4.5.2.1 Selection of Points

By default, the RMM selects random points and defines states over these points. Instead of the random selection, one may want to select more distinctive points from images. SIFT (scale invariant feature transform) is an algorithm that identifies keypoint locations on an image that are invariant with respect to image translation, scaling, and rotation. This keypoint localization is done by finding minima and maxima of the difference of Gaussian functions applied in scale space. Then some of these points are eliminated if they are on an edge or they have low contrast features [41]. In order to understand the effect of using the SIFT points in the RMM, we first compute the SIFT points for tissue images using the default parameters described in Lowe's paper [64]. In the perturbed sample generation step of the RMM, we select the data points randomly from the image pixels. To make use of the SIFT points in the RMM, we select these

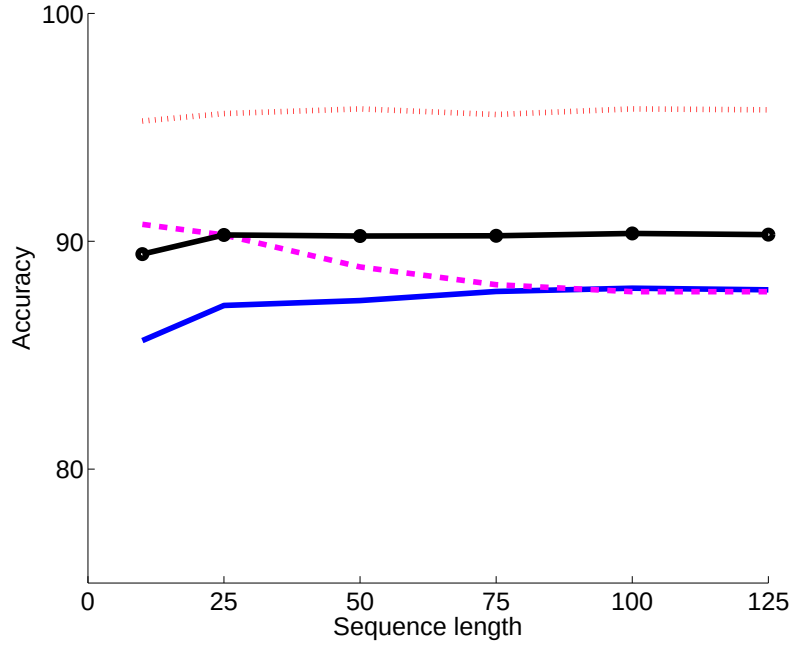


(a)

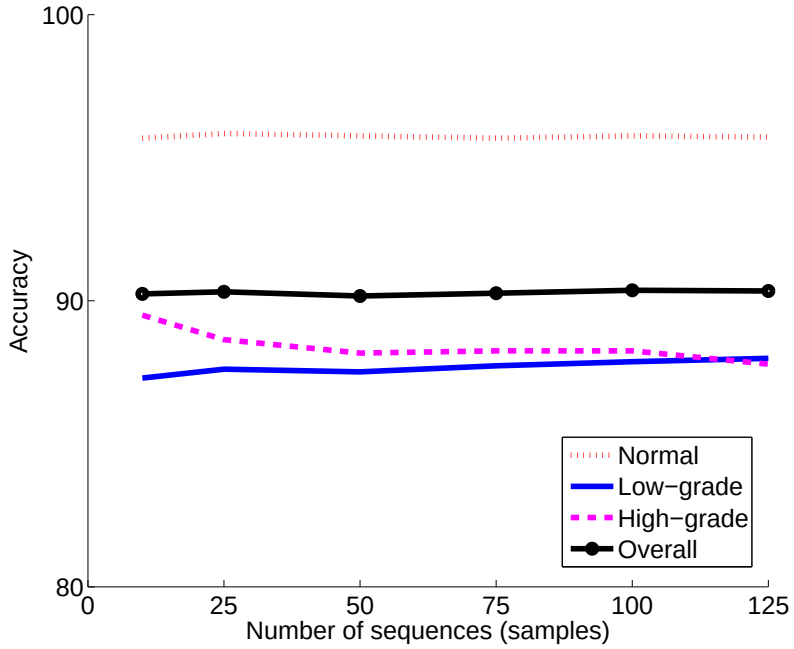


(b)

Figure 4.4: The test accuracies as a function of the model parameters: (a) window size, (b) number of states



(a)



(b)

Figure 4.5: The test accuracies as a function of the model parameters: (a) sequence length, (b) number of sequences.

data points randomly from the generated SIFT points for the image. The results when the RMM uses random points and the SIFT points are listed in Table 4.5. The Mc-Nemar test shows that there is not significant difference between these accuracies for $\alpha = 0.05$. The results show that the SIFT points do not carry additional information, compared to the random points in our application. However, it would be interesting to define a set of salient points by implementing a new algorithm that considers the domain specific knowledge (instead of using the SIFT algorithm, which does not make use of any domain specific knowledge) and select our points among the newly defined salient point set. We consider this as a future research aspect of the proposed method.

Table 4.5: Classification accuracies obtained by the RMM that uses random points and the RMM that uses SIFT points.

	Normal	Low	High	Overall
RMM with random points	95.64 ± 0.18	87.77 ± 0.32	88.56 ± 0.39	90.32 ± 0.18
RMM with SIFT points	96.04 ± 0.17	88.22 ± 0.27	88.96 ± 0.44	90.75 ± 0.14

4.5.2.2 Selected Feature Types

The RMM uses four features to characterize randomly selected points. Since the RMM does not require any specific feature type, one may use different types of features. To understand the success of the RMM with different features, we also repeat our experiments using the intensity (first-order statistics) features. The explicit parameters associated with this feature set are selected as the same with the $IntensityHistogramGrids$ features (see Section 4.2.2.2).

The test results of the RMM with the intensity features are given in Table 4.6. The RMM increases the overall accuracy of the $IntensityHistogramGrids$, which uses the same set of features, significantly. This supports the flexibility of RMM to be used with other feature types. The Mc-Nemar test indicates the difference between the RMM and *IntensityHistogramGrids* is statistically significant for $\alpha = 0.05$. Note that the accuracy of the RMM is lower than the one that uses the selected features. This indicates that with better features, the RMM gives higher

accuracies. On the other hand, the results given in Table 4.6 shows that when the same features are used, the RMM has a potential to increase the classification accuracy.

Table 4.6: Classification accuracies obtained by RMM, RMM with intensity and RMM with cooccurrence

	Normal	Low	High	Overall
RMM with Intensity Features	90.23 ± 0.45	85.42 ± 0.32	82.39 ± 0.69	86.41 ±0.20
IntensityHistogramGrids	78.82	74.17	78.60	76.32

4.5.2.3 Ordering Algorithm

After the RMM selects N points on an image, it orders them using a greedy heuristic that is an approximation of the Hamilton path (see Section 3.1). Figure 4.6(a) illustrates an example ordering for 100 random points using this heuristic. There are also different alternatives for ordering these points. For instance, the current approach orders these N points by selecting the most top-left point as an initial point. One alternative would be to use the same algorithm that starts ordering from a random point among these N points. Figure 4.6(b) plots an example ordering for 100 random points using this alternative. Another alternative would be to use Z-order to order these N points. Z-order is a space filling curve that maps multidimensional points into one dimensional data while preserving their locality. The Z value of points are calculated by interleaving the binary representation of their coordinates. Then the points are sorted according to their corresponding Z values to obtain their Z-order [44]. In Figure 4.6(c), we see an example of z-ordering on 100 given random points.. Another alternative would be employ Fiedler ordering. In order to calculate Fiedler ordering, one may construct a complete graph $G = (V, U)$ where V is the set of nodes, $U(X, Y)$ is a similarity function that gives the weight of the edge between $X \in V$ and $Y \in V$.

$$U(X, Y) = \frac{1}{||X - Y||} \tag{4.1}$$

Then, the Laplacian matrix $L(X,Y)$ of $G = (V,U)$ is calculated. The Laplacian matrix is equal to the difference of degree matrix and the adjacency matrix.

$$L(X,Y) = \begin{cases} -U(X,Y) & \text{if } X \neq Y \\ \sum_{Z \in V} U(X,Z) & \text{if } X = Y \end{cases} \tag{4.2}$$

Eigen value decomposition of the Laplacian matrix gives eigen values and corresponding eigen vectors. If we order these eigen values as $0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_N$, the eigen vector corresponding to second smallest eigen value λ_1 is named as the Fiedler vector. The ordering of the points, according to the order of corresponding sorted values in the Fiedler vector, gives Fiedler ordering [52]. Fiedler ordering is an approximation of the Hamilton path. Figure 4.6(d) illustrates the Fiedler ordering of given points.

Classification accuracies obtained by the RMM when these orderings are used reported in Table 4.7. The RMM with a greedy ordering approach is barely higher than the other three approaches. On the other hand, the Mc-Nemar test indicates that there is no significant statistical difference between these accuracies at $\alpha = 0.05$ level. This shows that the ordering of selected points with different approximations of the Hamilton path does not affect the performance of the RMM. However, one may work on developing ordering algorithms specifically designed for histopathological images. This may improve the classification results.

Table 4.7: Classification accuracies obtained by the RMM with a greedy heuristic with top-left start point, the RMM with a greedy heuristic with random start point, the RMM with Z ordering and the RMM with Fiedler Ordering

	Normal	Low	High	Overall
RMM with Top-Left Start	95.64 ± 0.18	87.77 ± 0.32	88.56 ± 0.39	90.32 ±0.18
RMM with Random Start	95.69 ± 0.18	87.70 ± 0.24	88.43 ± 0.44	90.28 ±0.17
RMM with Z Order	95.54 ± 0.18	87.25 ± 0.27	88.49 ± 0.35	90.01 ±0.14
RMM with Fiedler Order	95.26 ± 0.31	86.93 ± 0.44	89.73 ± 0.61	89.95 ±0.23

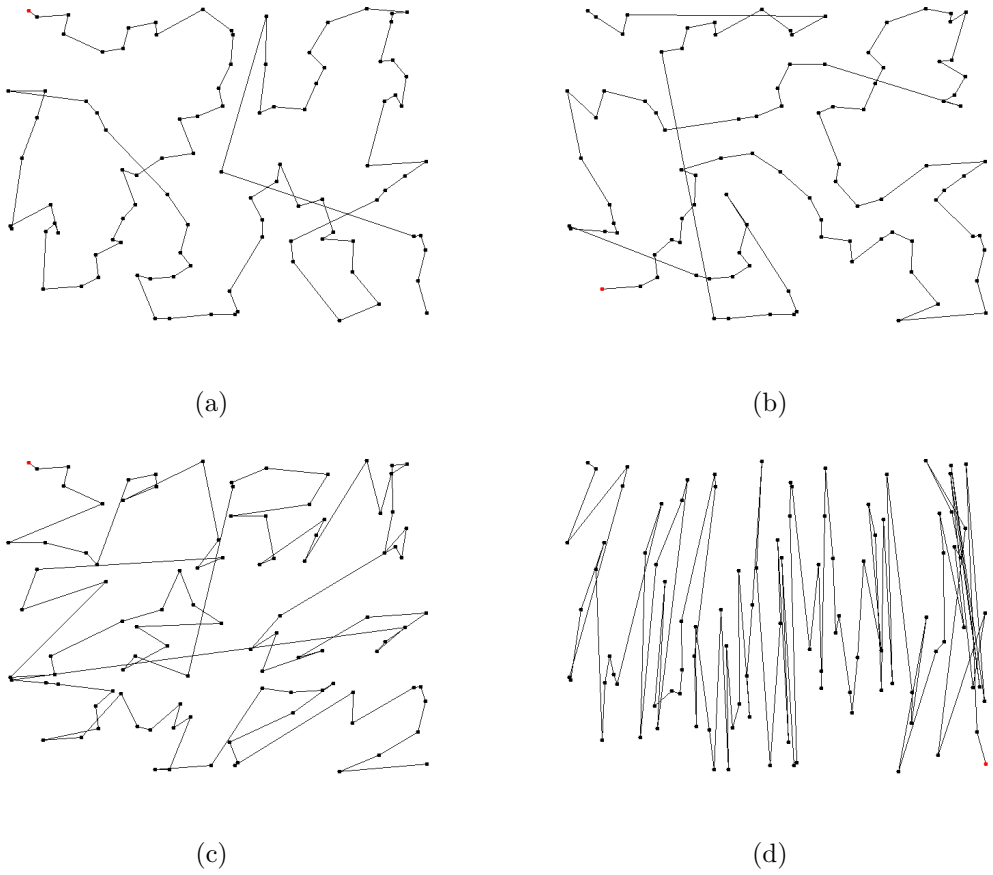


Figure 4.6: Different types of ordering algorithms for 100 random points: (a) greedy heuristic, (b) greedy random start, (c) Z-order, and (d) Fiedler order

4.5.2.4 Order of Markov models

A Markov model with an order of M assumes that the next state is dependent to M previous states. For example, in a second order Markov model, the next state is dependent to its two previous states. RMM uses first order Markov models. To understand the effect of this choice, we repeat our experiments holding other parameters fixed and using zero and second order Markov models instead of using first order Markov model. The results are listed in Table 4.8. The Mc-Nemar test indicates that the overall accuracy improvement of the first order Markov model over zero order Markov model is statistically significant for $\alpha = 0.05$. Between the second order Markov model and the first order Markov model, the

difference is not also statistically significant with $\alpha = 0.05$. Here, it is possible to use higher-order Markov models. Nevertheless, the use of higher order models 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 data. This may especially become a problem when the number of the training data is limited.

Table 4.8: Classification accuracies obtained by zero-order, first-order and second-order markov model

	Normal	Low	High	Overall
Zero-order MM	94.65 ± 0.38	82.71 ± 0.30	89.82 ± 0.59	87.54 ± 0.16
First-order MM	95.64 ± 0.18	87.77 ± 0.32	88.56 ± 0.39	90.32 ±0.18
Second-order MM	96.24 ± 0.24	89.74 ± 0.21	87.61 ± 0.46	91.40 ± 0.16

Chapter 5

Conclusion

This thesis 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 perturbed sample sequences from an image and models the samples 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 perturbed samples to classify an image, is also effective in increasing the classification accuracy.

As noted earlier, the proposed model does not impose any particular feature type to characterize selected data points. 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. As another alternative, one can 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 proposed model uses Markov processes for classification. It is also possible to use different classifiers (e.g., SVMs). In that case, instead of using sequences, a feature vector should be defined for an image using the features of its selected points. This would be another future research direction of this work.

Bibliography

[1] F. Albregtsen. Statistical texture measures computed from gray level run length matrices. *IMAGE*, 1(2):3–4, 1995.

[2] D. Altunbay, C. Cigir, C. Sokmensuer, and C. Gunduz-Demir. Color graphs for automated cancer diagnosis and grading. *IEEE Transactions on Biomedical Engineering*, 57(3):665–674, 2010.

[3] A. Andrion, C. Magnani, P. Betta, A. Donna, F. Mollo, M. Scelsi, P. Bernardi, M. Botta, and B. Terracini. Malignant mesothelioma of the pleura: interobserver variability. *Journal of Clinical Pathology*, 48(9):856, 1995.

[4] J. Arribas, V. Calhoun, and T. Adali. Automatic bayesian classification of healthy controls, bipolar disorder, and schizophrenia using intrinsic connectivity maps from fmri data. *IEEE Transactions on Biomedical Engineering*, 57(12):2850–2860, 2010.

[5] A. Basavanhally, S. Ganesan, S. Agner, J. Monaco, M. Feldman, J. Tomaszewski, G. Bhanot, and A. Madabhushi. Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. *IEEE Transactions on Biomedical Engineering*, 57(3):642–653, 2010.

[6] M. Bibbo, F. Michelassi, P. Bartels, H. Dytch, C. Bania, E. Lerma, and A. Montag. Karyometric marker features in normal-appearing glands adjacent to human colonic adenocarcinoma. *Cancer Research*, 50(1):147, 1990.

[7] C. Chang and C. Lin. Libsvm: a library for support vector machines. 2001.

[8] S. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, 1999.

[9] W. Chen, C. Metz, M. Giger, and K. Drukker. A novel hybrid linear/nonlinear classifier for two-class classification: Theory, algorithm, and applications. *IEEE Transactions on Medical Imaging*, 29(2):428–441, 2010.

[10] H. Choi, T. Jarkrans, E. Bengtsson, J. Vasko, K. Wester, P. Malmstrom, and C. Busch. Image analysis based grading of bladder carcinoma. comparison of object, texture and graph based methods and their reproducibility. *Analytical Cellular Pathology*, 15(1):1–18, 1997.

[11] A. Chu, C. Sehgal, and J. Greenleaf. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters*, 11(6):415–419, 1990.

[12] W. Cumberland and R. Royall. Does simple random sampling provide adequate balance? *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 118–124, 1988.

[13] L. Dalton, S. Pinder, C. Elston, I. Ellis, D. Page, W. Dupont, and R. Blamey. Histologic grading of breast cancer: linkage of patient outcome with level of pathologist agreement. *Modern Pathology*, 13(7):730–735, 2000.

[14] C. Demir, S. Gultekin, et al. Learning the topological properties of brain tumors. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 262–270, 2005.

[15] C. Demir, S. Gultekin, and B. Yener. Augmented cell-graphs for automated cancer diagnosis. *Bioinformatics*, 21(suppl 2):ii7, 2005.

[16] Y. Deng and B. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 800–810, 2001.

[17] 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 *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008.*, pages 496–499. IEEE, 2008.

[18] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi. A boosted bayesian multi-resolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE Transactions on Biomedical Engineering*, (99):1–1, 2010.

[19] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski. Automated grading of prostate cancer using architectural and textural image features. In *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*, pages 1284–1287. IEEE, 2007.

[20] R. Duda, P. Hart, D. Stork, et al. *Pattern classification*, volume 2. wiley New York, 2001.

[21] N. Esgiar, R. Naguib, B. Sharif, M. Bennett, and A. Murray. Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa. *IEEE Transactions on Information Technology in Biomedicine*, 2(3):197–203, 1998.

[22] A. Estabrooks, T. Jo, and N. Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36, 2004.

[23] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[24] A. Fischer, K. Jacobson, J. Rose, and R. Zeller. Hematoxylin and eosin staining of tissue and cell sections. *CSH protocols*, 2008:pdb–prot4986, 2008.

[25] M. Galloway. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*, 4(2):172–179, 1975.

[26] L. George and K. Sager. Breast cancer diagnosis using multi-fractal dimension spectra. In *IEEE International Conference on Signal Processing and Communications, 2007. ICSPC 2007.*, pages 592–595. IEEE.

[27] N. Ghoggali, F. Melgani, and Y. Bazi. A multiobjective genetic svm approach for classification problems with limited training samples. *IEEE Transactions on Geoscience and Remote Sensing*, 47(6):1707–1718, 2009.

[28] J. Gil, H. Wu, and B. Wang. Image analysis and morphometry in the diagnosis of breast cancer. *Microscopy Research and Technique*, 59(2):109–118, 2002.

[29] A. Graps. An introduction to wavelets. *IEEE Computational Science & Engineering*, 2(2):50–61, 1995.

[30] C. Gunduz-Demir. Mathematical modeling of the malignancy of cancer using graph evolution. *Mathematical Biosciences*, 209(2):514–527, 2007.

[31] R. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[32] P. Huang and C. Lee. Automatic classification for pathological prostate images based on fractal analysis. *IEEE Transactions on Medical Imaging*, 28(7):1037–1050, 2009.

[33] Q. Jackson and D. Landgrebe. An adaptive classifier design for high-dimensional data analysis with a limited training data set. *IEEE Transactions on Geoscience and Remote Sensing*, 39(12):2664–2679, 2001.

[34] A. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[35] A. Jemal, R. Siegel, J. Xu, and E. Ward. Cancer statistics, 2010. *CA: a Cancer Journal for Clinicians*, pages caac–20073v1, 2010.

[36] A. Karahaliou, I. Boniatis, S. Skiadopoulos, F. Sakellaropoulos, N. Arikidis, E. Likaki, G. Panayiotakis, and L. Costaridou. Breast cancer diagnosis: Analyzing texture of tissue surrounding microcalcifications. *IEEE Transactions on Information Technology in Biomedicine*, 12(6):731–738, 2008.

[37] G. Landini. *Applications of fractal geometry in pathology*. CRC Press, Boca Raton, FL, 1996.

[38] Y. Lee, L. Lee, and C. Tseng. Isolated mandarin syllable recognition with limited training data specially considering the effect of tones. *IEEE Transactions on Speech and Audio Processing*, 5(1):75–80, 1997.

[39] M. Li and Z. Zhou. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 37(6):1088–1098, 2007.

[40] Y. Liu. Active learning with support vector machine applied to gene expression data for cancer classification. *Journal of Chemical Information and Computer Sciences*, 44(6):1936–1941, 2004.

[41] D. Lowe. Object recognition from local scale-invariant features. In *iccv*, page 1150. Published by the IEEE Computer Society, 1999.

[42] M. Masud, J. Gao, L. Khan, J. Han, and B. Thuraisingham. A practical approach to classify evolving data streams: Training with limited amount of labeled data. In *Eighth IEEE International Conference on Data Mining, 2008. ICDM'08.*, pages 929–934. IEEE, 2008.

[43] J. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, A. Glass, B. Zehnbauer, K. Lister, and R. Parwaresch. Breast carcinoma malignancy grading by bloom–richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. *Modern pathology*, 18(8):1067–1078, 2005.

[44] G. Morton. *A computer oriented geodetic data base and a new technique in file sequencing*. International Business Machines Co., 1966.

[45] G. Morton. Resampling methods for unsupervised learning from sample data, machine learning. 2009.

[46] K. Nazarpour, C. Ethier, L. Paninski, J. Rebesco, R. Miall, and L. Miller. Emg prediction from motor cortical recordings via a non-negative point process filter. *IEEE Transactions on Biomedical Engineering*, (99):1–1, 2011.

[47] A. Nickerson, N. Japkowicz, and E. Milios. Using unsupervised learning to guide re-sampling in imbalanced data sets. In *Proceedings of the Eighth International Workshop on AI and Statitsics*, pages 261–265. Citeseer, 2001.

[48] T. Ojala and M. T. Pietikainen, M. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, pages 971–987, 2002.

[49] A. Özçift. Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. *Computers in Biology and Medicine*, 2011.

[50] D. Padfield, B. Chen, H. Roysam, C. Cline, G. Lin, and M. Seel. Cancer tissue classification using nuclear feature measurements from dapi stained images. In *Proc. 1st Workshop Microscopic Image Anal. Appl. Biol*, pages 86–92, 2006.

[51] H. Qureshi, O. Sertel, N. Rajpoot, R. Wilson, and M. Gurcan. Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008*, pages 196–204, 2008.

[52] A. Robles-Kelly and E. Hancock. Graph edit distance from spectral seriation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 365–378, 2005.

[53] B. Sahiner, H. Chan, and L. Hadjiiski. Classifier performance prediction for computer-aided diagnosis using a limited dataset. *Medical Physics*, 35:1559, 2008.

[54] O. Sertel, J. Kong, H. Shimada, U. Catalyurek, J. Saltz, and M. Gurcan. Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. *Pattern Recognition*, 42(6):1093–1103, 2009.

[55] B. Settles. Active learning literature survey. *Science*, 10(3):237–304, 1995.

[56] H. Soltanian-Zadeh and K. Jafari-Khouzani. Multiwavelet grading of prostate pathological images.

[57] W. Street, W. Wolberg, and O. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Proceedings of SPIE*, volume 861, 1905.

[58] J. Sudbo, R. Marcelpoil, and A. Reith. New algorithms based on the voronoi diagram applied in a pilot study on normal mucosa and carcinomas. *Analytical Cellular Pathology*, 21(2):71–86, 2000.

[59] A. Tabesh, M. Teverovskiy, H. Pang, V. Kumar, D. Verbel, A. Kotsianti, and O. Saidi. Multifeature prostate cancer diagnosis and gleason grading of histological images. *IEEE Transactions on Medical Imaging*, 26(10):1366–1378, 2007.

[60] J. Thiran and B. Macq. Morphological feature extraction for the classification of digital images of cancerous tissues. *IEEE Transactions on Biomedical Engineering*, 43(10):1011–1020, 1996.

[61] A. Tosun, M. Kandemir, C. Sokmensuer, and C. Gunduz-Demir. Object-oriented texture analysis for the unsupervised segmentation of biopsy images for cancer detection. *Pattern Recognition*, 42(6):1104–1112, 2009.

[62] G. Tourassi. Journey toward computer-aided diagnosis: Role of image texture analysis1. *Radiology*, 213(2):317, 1999.

[63] G. Tourassi and C. Floyd. The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis. *Medical Decision Making*, 17(2):186, 1997.

[64] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

[65] W. Wang, J. Ozolek, and G. Rohde. Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images. *Cytometry Part A*, 77(5):485–494, 2010.

[66] B. Weyn, W. Jacob, V. da Silva, R. Montironi, P. Hamilton, D. Thompson, H. Bartels, A. Van Daele, K. Dillon, and P. Bartels. Data representation and reduction for chromatin texture in nuclei from premalignant prostatic, esophageal, and colonic lesions. *Cytometry*, 41(2):133–138, 2000.

[67] B. Weyn, G. Van De Wouwer, 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 mesothelioma. *The Journal of Pathology*, 189(4):581–589, 1999.

[68] 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*, 35(1):23–29, 1999.

[69] M. Wiltgen, A. Gerger, and J. Smolle. Tissue counter analysis of benign common nevi and malignant melanoma. *International Journal of Medical Informatics*, 69(1):17–28, 2003.

[70] H. Yu, G. Gu, H. Liu, J. Shen, C. Zhu, and J. Ni. Incremental tumor diagnosis algorithm using new unlabeled microarray. In *2008 International Multi-symposiums on Computer and Computational Sciences*, pages 45–50. IEEE, 2008.

[71] J. Ziv. An efficient universal prediction algorithm for unknown sources with limited training data. *IEEE Transactions on Information Theory*, 48(6):1690–1693, 2002.