COLOR GRAPH REPRESENTATION FOR STRUCTURAL ANALYSIS OF TISSUE IMAGES

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Doğan Altunbay

July, 2010

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.

Assist. Prof. Dr. Selim Aksoy

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. Tolga Can

Approved for the Institute of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Institute

ABSTRACT

COLOR GRAPH REPRESENTATION FOR STRUCTURAL ANALYSIS OF TISSUE IMAGES

Doğan Altunbay
M.S. in Computer Engineering
Supervisor: Assist. Prof. Dr. Çiğdem Gündüz Demir
July, 2010

Computer aided image analysis tools are becoming increasingly important in automated cancer diagnosis and grading. They have the potential of assisting pathologists in histopathological examination of tissues, which may lead to a considerable amount of subjectivity. These analysis tools help reduce the subjectivity, providing quantitative information about tissues. In literature, it has been proposed to implement such computational tools using different methods that represent a tissue with different set of image features. One of the most commonly used methods is the structural method that represents a tissue quantifying the spatial relationship of its components. Although previous structural methods lead to promising results for different tissue types, they only use the spatial relations of nuclear tissue components without considering the existence of different components in a tissue. However, additional information that could be obtained from other components of the tissue has an importance in better representing the tissue, and thus, in making more reliable decisions.

This thesis introduces a novel structural method to quantify histopathological images for automated cancer diagnosis and grading. Unlike the previous structural methods, it proposes to represent a tissue considering the spatial distribution of different tissue components. To this end, it constructs a graph on multiple tissue components and colors its edges depending on the component types of their end points. Subsequently, a new set of structural features is extracted from these "color graphs" and used in the classification of tissues. Experiments conducted on 3236 photomicrographs of colon tissues that are taken from 258 different patients demonstrate that the color graph approach leads to 94.89 percent training accuracy and 88.63 percent test accuracy. Our experiments also show that the introduction of color edges to represent the spatial relationship of different tissue components and the use of graph features defined on these color edges significantly improve the classification accuracy of the previous structural methods.

Keywords: medical image analysis, graph theory, histopathological image analysis, cancer diagnosis and grading.

ÖZET

DOKU İMGELERİNİN YAPISAL ANALİZİ İÇİN RENKLİ ÇİZGE GÖSTERİMİ

Doğan Altunbay
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Yar. Doç. Dr. Çiğdem Gündüz Demir
Temmuz, 2010

Bilgisayar destekli görüntü analizi araçları, otomatikleştirilmiş kanser tanı ve derecelendirmesi alanında giderek önemli hale gelmektedir. Bu araçların, kayda değer öznelliklere neden olabilen doku histopatolojik incelemesinde patologlara yardımcı olma potansiyelleri bulunmaktadır. Bu analiz araçları, dokular ile ilgili nicel bilgi sağlayarak öznelliğin azaltılmasına yardımcı olmaktadır. Literatürde, bu tip hesaplamasal araçların, dokuyu farklı imge öznitelikleri ile gösteren farklı yöntemler kullanarak geliştirilmesi önerilmiştir. En çok kullanılan yöntemlerden biri, doku bileşenleri arasındaki uzamsal ilişkiyi niceleyerek dokuyu temsil eden yapısal yöntemdir. Önceki yapısal yöntemler ile değişik doku türleri için umut verici sonuçlar elde edilmesine rağmen, bu yöntemler, bir dokunun nicelenmesi için yalnızca çekirdek bileşenlerini kullanmakta ve dokudaki diğer bileşenlerin varlığını dikkate almamaktadır. Öte yandan, bir dokunun değişik bileşenlerinden elde edilebilecek ek bilgi, bu dokunun daha iyi gösterilmesinde, ve dolayısıyla daha güvenilir kararlar alınmasında önemlidir.

Bu tez, otomatik kanser tanı ve derecelendirmesi için histopatolojik imgelerin nicelenmesinde kullanılabilecek yeni bir yapısal yöntem sunmaktadır. Önceki yöntemlerin aksine, önerilen yöntem, farklı doku bileşenlerinin uzamsal dağılımlarını dikkate alarak dokuyu temsil etmeyi önermektedir. Bu amaçla, önerilen yöntem farklı doku bileşenlerin üzerinde bir çizge tanımlar ve bu çizgenin kenarlarını uç bileşenlerinin türlerine göre renklendirir. Ardından, elde edilen bu “renkli çizgeler”den yeni bir yapısal öznitelik kümesi çıkarır ve bu kümeyi dokuların sınıflandırılmasında kullanır. 258 farklı hastadan alınan 3236 kolon doku imgesi üzerinde yapılan deneyler, önerilen renkli çizge yönteminin, öğrenme seti için yüzde 94.89 ve test seti için yüzde 88.63 doğruluk verdiğini göstermektedir. Deneylerimiz ayrıca, farklı doku bileşenleri arasındaki uzamsal ilişkiyi gösteren renkli kenarların tanımlanmasının ve bunlar üzerinde çıkarılan çizge özniteliklerinin kullanımının, önceki yapısal yöntemlerin sınıflandırma başarısını kayda değer ölçüde artırdığını göstermektedir.

Anahtar sözcükler: Medikal görüntü analizi, çizge teorisi, histopatolojik görüntü analizi, kanser tanı ve derecelendirilmesi.

Acknowledgement

I present my thanks and respect to my advisor Assist. Prof. Dr. Çiğdem Gündüz Demir for her motivating guidance throughout this study. I also thank Assist. Prof. Dr. Selim Aksoy and Assist. Prof. Dr. Tolga Can for accepting to participate in my thesis committee.

I am grateful to my parents, Gülşen and Hakim Altunbay, my brothers, Cihan and Hakan Can Altunbay, and my fiance, Büşra Aşlar, for their endless support.

Doğan Altunbay
July 2010

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer is a member of neoplastic diseases. It may occur in various tissue types and in different forms, and causes abnormal growth and change in the cellular structure of tissues from which it originates. With the increased malignancy level of cancer, tissues eventually lose their distinguishing characteristics. This situation may result in the interruption of the vital functions of organs, which makes cancer one of the most lethal diseases [1, 46, 63]. However, with its early detection and the correct treatment selection, for which cancer grade is an important factor, survival rates greatly increase. A considerable amount of effort has been put in the field of cancer diagnosis and grading. Medical experts take advantage of numerous medical imaging techniques such as Magnetic Resonance Imaging (MRI), Computer Aided Tomography (CAT), Mammography, Colonoscopy, Ultrasound Imaging for cancer screening. Although these methods provide effective diagnosis tools for screening and early detection of tumors, they may not be helpful in determining their malignancy level. Moreover, these methods are not used as the gold standard and biopsy examination is still necessary to reach the final decision.

In the current practice of medicine, histopathological examination is the routinely used method to examine biopsies for cancer diagnosis and grading [40, 68]. In this examination, a biopsy specimen is surgically removed from the patient and prepared following suitable fixation and staining procedures. Afterwards, a pathologist visually inspects the biopsy under a microscope and determines how much the tissue differentiates from its normal appearance. The pathologist then makes a decision about the existence of cancer, and its grade depending on his/her observation.

The extent of differentiation in the tissue, which is determined by histopathological examination, serves as a basis in the choice of relevant medical and/or surgical treatment method. Early detection and correct grading of cancer affect the success of the selected treatment method and increase the chance of survival [2]. Therefore, it is important to use procedures that provide efficient information about the malignancy of tissues. Although histopathological examination provides efficient information for accurate diagnosis and grading [59, 64, 14, 20], it is subject to a considerable amount of subjectivity as it is mainly based on visual interpretation of pathologists, which also requires expertise in the field. Based on their experience and interpretation, different pathologists may come up with different decisions for the same tissue. Moreover, a pathologist may make different decisions for the same tissue at different times. This inter and intra observer variability may reach significant levels [93, 33] and directly affects the accuracy of the procedure.

Such observer variability reveals a necessity for employing quantitative information to assist the pathologists during the diagnosis and grading processes. Many studies have proposed computational methods in order to obtain quantitative information about tissues. These methods are used for developing computerized diagnosis tools that provide estimates of the malignancy level and help alleviate the subjectivity of the pathologists.

1.1 Motivation

There are numerous studies that focus on the computer aided diagnosis and grading of cancer. These studies propose image analysis frameworks that employ computational methods for the quantification of tissues. These frameworks make use of morphological, textural, and structural properties of histopathological images to represent a tissue for its quantification.

The morphological approach relies on the use of shape and size characteristics of cellular components. For this purpose, it extracts features such as area, perimeter, roundness, and symmetry for the quantification of a tissue [106, 86, 88]. The computation of these features transforms pixel level information to component level information. On the other hand, this computation requires segmenting the tissue in order to identify exact boundaries of its components. Complex structure of a typical histopathological image increases the difficulty in segmentation, and may necessitate the manual segmentation of the tissue.

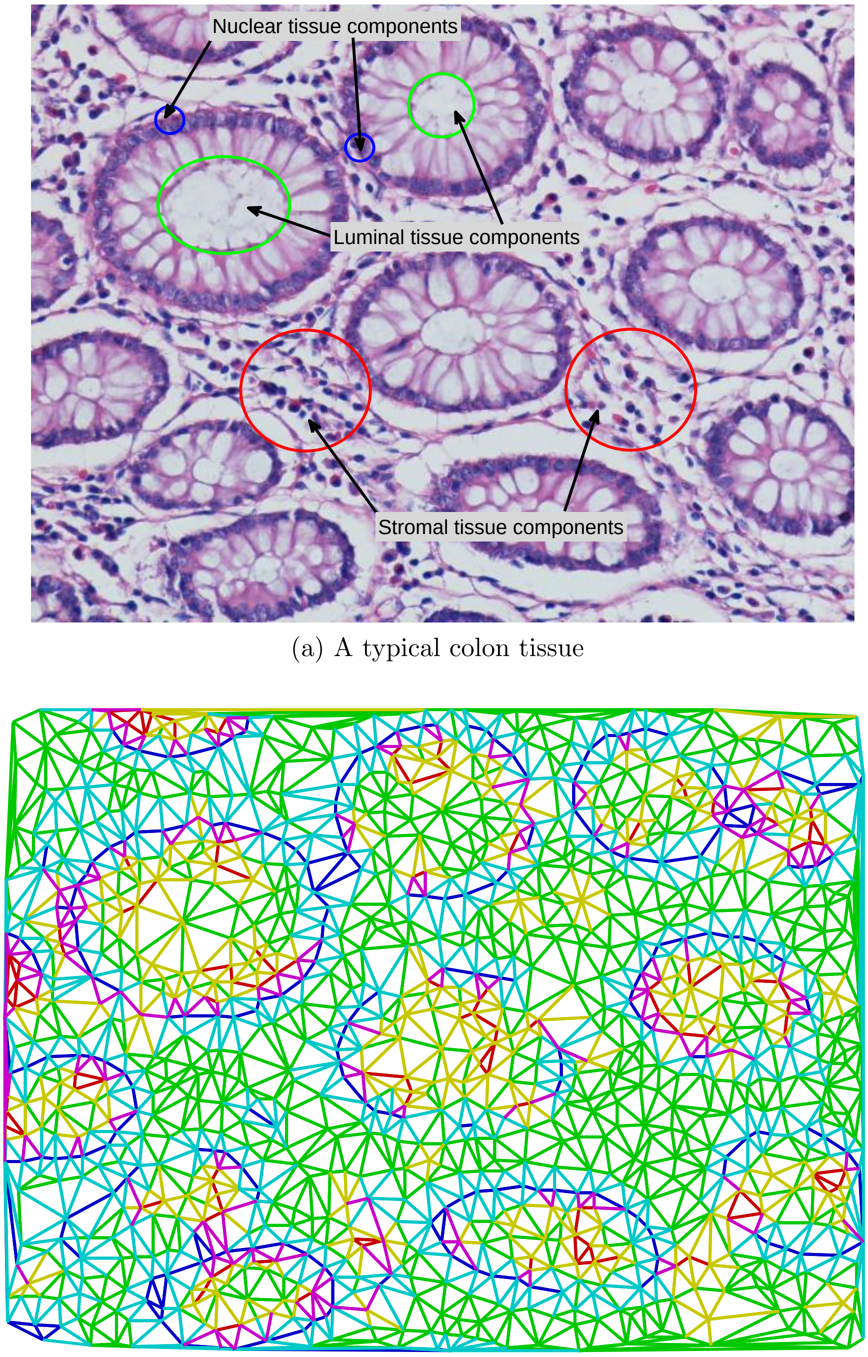
In the textural approach, a tissue is represented with texture characteristics of either the whole tissue [37, 38, 82] or particular components of the tissue such as nuclei [19, 103, 101, 48]. The studies make use of textural features that are computed from co-occurrence matrices [37, 82, 38], run length matrices [104, 4], multiwavelet coefficients [101, 27, 62], and fractal geometry [8, 34, 39]. The definition of these features primarily depends on the pixels of tissue images, thus they are sensitive to noise in the pixel values. Besides, there could be a large amount of noise in a typical histopathological image due to the problems occurred in staining and sectioning [48]. Statistical features computed from histograms of pixel intensities such as average, standard deviation, and entropy [105], and optical densities of nuclear tissue components are also used [104] for quantifying the texture characteristics of a tissue. However, histopathological images of different tissue types usually have similar intensity distributions as they are stained using the same procedure, and this introduces an important drawback for these methods.

The structural approach provides a higher level representation of a tissue using the spatial relationships and neighborhood properties of their cellular components. To this end, a graph is constructed on the cellular components and a set of local and global graph features is extracted [37, 104]. In literature, there are different methods of generating graphs. The most common method is to use Delaunay triangulations (and their corresponding Voronoi diagrams) where nuclear components of the tissue are considered as graph nodes [102, 65, 90, 52, 36, 17]. Minimum spanning trees obtained from these graph based representations are also used for representation. Another method is to use probabilistic graphs in which nuclear components are considered as graph nodes, and edges are probabilistically assigned between these nodes [53, 30, 29, 11].

1.2 Contribution

The aforementioned structural methods provide powerful frameworks for computerized cancer diagnosis and grading. On the other hand, these methods consider nuclear components of tissues, ignoring the existence of other tissue components such as luminal and stromal regions. Nevertheless, information extracted also considering these additional tissue components may improve the effectiveness of the representation. This improvement can be significant especially for tissues where the components are hierarchically organized. Colon tissues are the examples of such tissues. They are formed of nuclear, luminal, and stromal tissue components. Nuclei of the epithelial cells of colon tissue are lined up around its luminal components and form glandular structures of the colon. Stromal tissue components are distributed between these glandular regions. This hierarchical organization of its nuclear, luminal, and stromal components reflects the major characteristics of the colon tissue. Figure 1.1a shows the histopathological tissue image of a colon tissue section. As observed in this figure, in addition to nuclear components, the distribution of luminal and stromal tissue components has an important role in describing the colon tissue structure. Pathologists also take advantage of spatial distributions of these tissue components in histopathological examination. Therefore, considering additional information obtained from these components in a structural representation will result in a better tissue quantification.

This thesis introduces a novel structural representation in which the spatial relationships of nuclear, stromal, and luminal tissue components are considered [5]. This new representation is used for automated diagnosis and grading of cancer.


Nuclear tissue components
Luminal tissue components
Stromal tissue components

(a) A typical colon tissue

(b) Color graph representation of the tissue image shown in (a)

Figure 1.1: Histological image of a typical colon tissue and the corresponding color graph representation.

The proposed representation first segments a tissue image into its components, transforming image pixels into circular primitives with the use of a heuristic approach described in [94, 55]. Considering these components as nodes, a graph is constructed using Delaunay triangulation. Afterwards, each edge of the triangulation is assigned a color label based on the types of its end nodes. Figure 1.1b shows the corresponding color graph representation of the tissue image given in Figure 1.1a. A set of new features is introduced for the constructed "color graph", which quantifies the relationship of different types of graph nodes (tissue components), and used for classification.

Experiments conducted on 3236 images taken from 258 patients have shown that color graph method yields accurate results for a three class classification problem on colon cancer, and outperforms previous structural methods. We have also shown that introduction of new features that quantify spatial relationships and neighborhood characteristics of different tissue components improves the discriminative power of graph based representation.

1.3 Outline of Thesis

The outline of this thesis is as follows. Chapter 2 provides background information about the problem domain, and summarizes previous computational methods on automated cancer diagnosis and grading. Chapter 3 provides detailed descriptions of the proposed color graph representation and newly introduced graph features. Chapter 4 gives the experimental setup and reports the results of automated colon cancer diagnosis and grading. Additionally, Chapter 4 gives the evaluation of the color graph method and its comparisons with other structural methods. Chapter 5 includes concluding remarks and discussions.

Chapter 2

Background

This chapter provides background information about the studies in the field of computer aided cancer diagnosis and grading. In the first section, the structure of a colon tissue is described and the information of colon cancer, which is the cancer type of interest in this thesis, is given. In the second section, image analysis methods that are employed by the previous studies on computer aided cancer diagnosis and grading are explained.

2.1 Domain Description

Gastrointestinal system of the human body is responsible for processing food for energy and excreting the solid waste out of the body. Colon is the first and the longest part of the large intestine and is formed of epithelial cells. The distinctive characteristics of colon epithelium is that it contains a large number of glandular structures. Glands in the colon tissue are responsible for two operations; absorbing water and mineral nutrients from the feces back into the blood and secreting mucus into the colon lumen for lubricating the dehydrated feces.

Colon cancer is a tumor forming type of cancer and it develops in the colon epithelium. It is the third most common cancer incidence among both men and women in Northern America [1]. Development of colon cancer is usually spread out over a time period of years. It usually begins as a noncancerous polyp, a growth of epithelial tissue that occurs on the lining of the colon, which may later change into cancer. *Adenomatous polyps* or *adenomas* are the kind of polyps which have glandular origin and are most likely to become cancerous.

According to the studies, more than half of all individuals will eventually develop one or more adenomas [81]. Although most adenomas are benign which do not turn into cancer, a considerable amount of them are malignant and turn into adenocarcinomas which account for 96 percent of colon cancers [87]. When it turns into cancer, an adenocarcinoma can grow through the lining and into the colon. With this unexpected growth of the adenocarcinoma, cancerous cells may spread into lymph vessels and blood vessels. Cancerous cells may also be carried in the lymph vessels and blood vessels to the other body parts such as lung and liver causing different types of cancer incidents. This process through which the cancerous cells spread to distant body parts is called *metastasis*.

In medicine, a wide range of methods are employed in the diagnosis of colon adenocarcinoma. Screening methods such as fecal occult blood test, fecal immunochemical test, and stool DNA test are applied for examining the existence of cancerous cells to detect the disease in its early stages. On the other hand, methods such as flexible sigmoidoscopy, colonoscopy and CT colonography allow structural examination of the colon tissue and help not only detection of adenomas, which are associated with an increased risk of cancer, but also removal of them [85, 2].

These methods provide comprehensive information for the diagnosis and early detection of cancerous cells in the colon. However they do not provide information about the degree of malignancy of the adenomas. For this reason, histopathological examination, which is a microscopic level inspection procedure, is conventionally used by pathologists in diagnosis and grading of colon adenocarcinoma. This procedure is also applied in the diagnosis and grading of a wide range of cancer types.

In this procedure, a specimen, which is considered as the main subject throughout the process, is extracted from the tissue making use of a surgical procedure called *biopsy*. Next, sections are cut from the biopsy specimen and stained with a chemical material for microscopic inspection. Staining is a chemical reaction that is used to increase the contrast in microscopic images of biopsy specimens allowing a better visual perception of different structures of the tissue.

Figure 2.1 shows a histological image of a section from a colon tissue, which is stained with the routinely used hematoxylin-and-eosin technique [40, 68]. In this figure, the components and the glandular structure of the colon tissue is indicated. An epithelial cell, marked with green circle in Figure 2.1, is formed of a nucleus and cytoplasm which correspond to dark purple and white colored areas in the image respectively. Epithelial cells in the colon are lined up around a vacant region, called luminal area or *lumen*, forming the *gland* structure. Absorption of water and nutrients and secretion of mucus is performed in lumens. Lumens also correspond to white colored regions in Figure 2.1. The other components in the tissue are stroma, which corresponds to pink colored regions in the figure. These components are responsible for holding the tissue together. There are also cells in stromal region. The nuclei of these cells are also shown in dark purple and their cytoplasm are usually not observed.

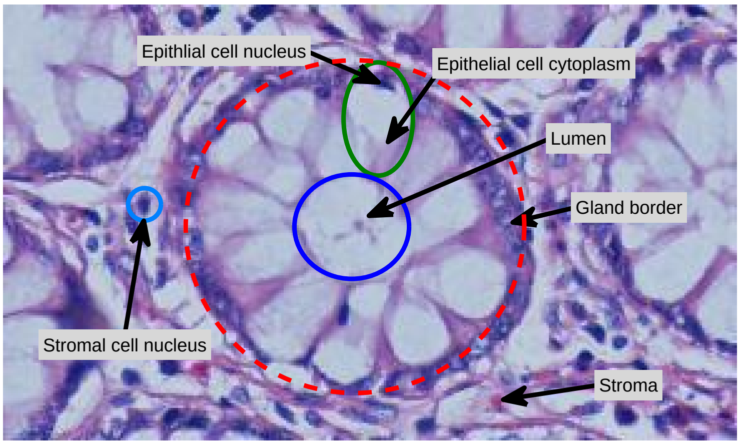
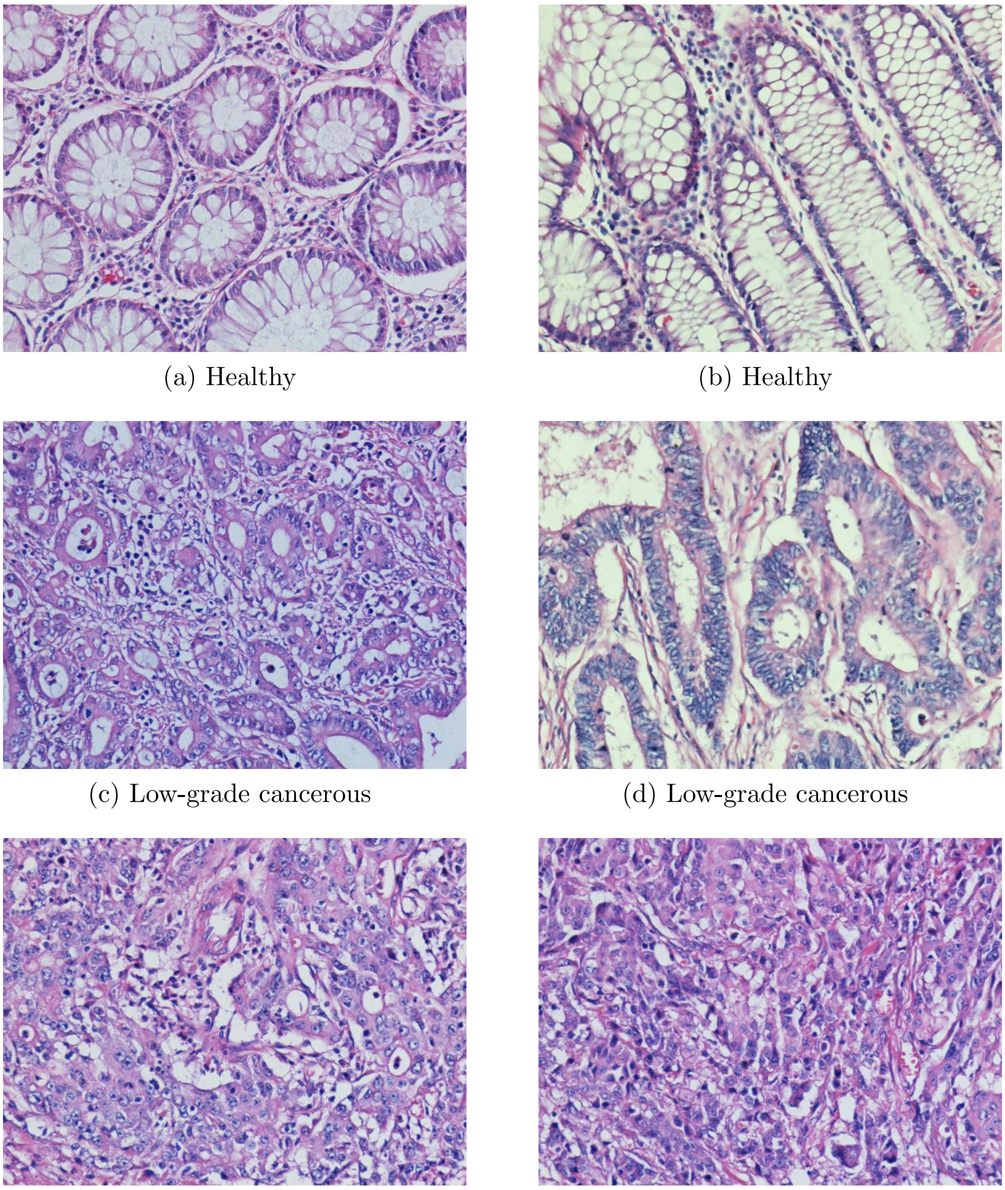


Figure 2.1: Microscopic view of a colon tissue section.

In a typical healthy colon tissue, there exist a large number of glands. Figure 2.2a and Figure 2.2b show histopathological images of this kind of healthy colon tissues under microscope. The glandular structures indicated in Figure 2.1



(a) Healthy

(b) Healthy

(c) Low-grade cancerous

(d) Low-grade cancerous

(e) High-grade cancerous

(f) High-grade cancerous

Figure 2.2: Histopathological images of colon tissues. These tissues are stained with the hematoxylin-and-eosin technique, which is a routinely used staining procedure in histopathology.

are easily perceived in these images. In the case of colon adenocarcinoma, these glandular structures are deformed, and eventually destroyed as the disease develops. In the inception phase of the development of cancer, the spatial distribution of epithelial cells looks similar to those of a normal tissue, where the glandular structures are well to moderately differentiated. Figure 2.2c and Figure 2.2d show sample images from this kind of colon tissues, which are classified as low grade cancerous. Further development of cancer introduces a higher level of distortion to the tissue, causing the tissue components turn into a poorly differentiated distribution and glandular structures totally disappear. Figure 2.2e and Figure 2.2f show sample images of this kind of colon tissues, which are classified as high grade cancerous.

A pathologist visually inspects biopsy specimens to determine the existence of malignant tumors in a tissue and their grades. Histopathological examination provides valuable information for the pathologist [14], however, the accuracy of his/her decisions significantly relies on his/her expertise. Based on the experience, histopathological examination may introduce considerable degrees of inter- and intra-observer variability. Besides, this variability may affect the selection of an appropriate treatment method.

2.2 Computer Aided Diagnosis and Grading of Cancer

The observer variability involved in histopathological examination reveals the need of quantitative methods for cancer diagnosis and grading that reduce the subjectivity of the experts. In literature, numerous studies have been proposed to provide computer aided image analysis frameworks that make use of mathematical models for tissue representation. Computational methods in this context can be grouped into three; morphological, textural, and structural. This section provides information about these computational methods.

2.2.1 Morphological Methods

Cancer also causes changes in the geometric structure of cellular components. The degree of this change may indicate the malignancy of tumor. Morphological approach relies on the shape and size characteristics of its nuclear components to represent a tissue [104, 48, 88, 86]. To this end, local features of individual cell nuclei such as area, perimeter, roundness, and compactness are defined. To describe the whole tissue, the average and standard deviation of these local features are computed.

Morphological approach is employed for the diagnosis and grading of different types of cancer. Street et al. [88] and Wolberg et al. [106] made use of morphological features of breast cell nuclei to obtain quantitative information about the breast tissue. Guillaud [52] also proposed to use nuclear morphometry in the diagnosis of cervical intraepithelial neoplasia. Such morphological featuers are also used to quantify different tissue components. For example, Doyle et al. [37] focused on morphological features of glandular structures in a prostate tissue to achieve an automated diagnosis framework. Anderson et al. [6] also made use of morphological features of glandular tissue components for discriminating malignant and benign tumors in a breast tissue.

2.2.2 Textural Methods

Texture analysis is a widely used method in image processing applications. The aim of this approach is to extract distinguishing and descriptive features from images for characterizing their textures. These features include first order statistical features such as mean and variance, which are related to values of individual pixels, and second order statistical features such as correlation and contrast which account for co-occurrence or inter-dependency of pixel pairs.

2.2.2.1 Intensity Based Features

Intensity based features are used for describing pixel level characteristics of images. To this end, gray level or color histograms of intensity values and densitometric features are employed. Properties of these features such as mean, standard deviation, kurtosis, and skewness are computed to obtain first order statistical information about the texture of tissues.

In literature, there have been less number of studies that employ intensity based image features for the automated diagnosis and grading of cancer. Nevertheless, these features are used together with other image features to improve quantitative power of proposed representations. Wiltgen [105] used gray level histograms in combination with co-occurrence features for classification of benign common nevi and malignant melanoma. Weyn et al. [101] employed optical densities of nuclear tissue components in the quantification of breast tissue images. The same group [104] used features obtained from densitometric properties of nuclear tissue components together with morphological and structural image features in the diagnosis and prognosis of malignant mesothelioma.

2.2.2.2 Co-occurrence Matrices

One of the most commonly used methods in texture analysis is co-occurrence matrices. Haralick [58] first proposed to use gray level co-occurrence matrices for the definition of textural image features. The values of a co-occurrence matrix represent the frequencies of pixels occurring in a relative distance to each other, one having the intensity value i and the other having the intensity value j. For a gray level image I with a size of $w \times h$, a co-occurrence matrix M with relative distance $d\left(x,y\right)$ is defined as follows:

$$M\left(i,j\right)_{d(x,y)} = \sum_{p=1}^{w}\sum_{q=1}^{h}\begin{cases} 1, & \text{if } \; I\left(p,q\right) = i \text{ and } I\left(p+x,q+y\right) = j \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

Angular second moment	$f_1 = \sum_i \sum_j M(i,j)^2$
Contrast	$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \begin{array}{l} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} M(i,j) \\ \lvert i-j \rvert = n \end{array} \right\}$
Correlation	$f_3 = \frac{\sum_i \sum_j (ij) M(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$, where μ_x , μ_y, σ_x, and σ_y are the means and standard deviations of M_x and M_y the partial probability density functions
Variance	$f_4 = \sum_i \sum_j (i-\mu)^2 M(i,j)$
Inverse difference moment	$f_5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} M(i,j)$
Sum average	$f_6 = \sum_{i=2}^{2N_g} i M_{x+y}(i)$, where x and y are the row and column numbers of an entry in the co-occurrence matrix, and $M_{x+y}(i)$ is the probability of co-occurrence matrix indices summing to $x+y$
Sum variance	$f_7 = \sum_{i=2}^{2N_g} (i - f_8)^2 M_{x+y}(i)$
Sum entropy	$f_8 = -\sum_{i=2}^{2N_g} M_{x+y}(i) \, log \{M_{x+y}(i)\}$
Entropy	$f_9 = -\sum_i \sum_j M(i,j) \, log \{M(i,j)\}$
Difference variance	$f_{10} = \sum_{i=0}^{N_g-1} i^2 M_{x-y}(i)$
Difference entropy	$f_{11} = -\sum_{i=0}^{N_g-1} M_{x-y}(i) \, log \{M_{x-y}(i)\}$
Information measures of correlation	$f_{12} = \frac{hxy - hxy1}{max\{hx, hy\}}$ $f_{13} = (1 - exp[-2(hxy2 - hxy)])^{1/2}$ $hxy = -\sum_i \sum_j M(i,j) \, log(M(i,j))$, where hx and hy are entropies of M_x and M_y, and $hxy1 = -\sum_i \sum_j M(i,j) \, log \{M_x(i) M_y(j)\}$ $hxy2 = -\sum_i \sum_j M_x(i) M_y(j) \, log \{M_x(i) M_y(j)\}$
Maximal correlation coefficient	$f_{14} =$ (Second largest eigenvalue of $Q)^{1/2}$ where $Q(i,j) = \sum_k \frac{M(i,k) M(j,k)}{M_x(i) M_y(j)}$

Table 2.1: Haralick features obtained from gray level co-occurrence matrices.

The entries of co-occurrence matrices are not usually used directly as image features. Instead, a number of representative features are extracted from these co-occurrence matrices. Table 2.1 lists the definitions of the co-occurrence matrix features proposed by Haralick [58].

In general, co-occurrence matrices are computed from gray level intensities as first described by Haralick. However, in literature, there are a number of studies that use co-occurrence matrices computed from color images. Features extracted from these color co-occurrence matrices are also used for representing the image texture [83, 84]. Co-occurrence matrices are sensitive to rotation, and thus, they are generally computed using varying offset sets, such as 0, $\pi/4$, $\pi/2$, $3\pi/4$, and the resulting co-occurrence matrices are combined to achieve rotation invariance [17, 38, 82].

Co-occurrence matrices are also proposed to be used in quantifying histopathological images of different types of tissues [57, 102, 104, 17, 37]. Esgiar et al. [38] used co-occurrence features for automated categorization of normal and cancerous colonic mucosa. Wiltgen et al. [105] employed co-occurrence matrices to describe texture characteristics of histopathological images of skin tissues. Co-occurrence features have also been employed for the quantification of chromatin texture of cell nuclei [99, 101, 103].

2.2.2.3 Gray Level Run Length Matrices

Gray level run length method, defined by Galloway [45], is another textural approach that provides higher order statistical information [92]. An entry $M(i, j)$ of a gray level run length matrix M is defined as the number of sequences of pixels that have the gray level intensity value i and are of length j. Galloway first defined five statistical features on gray level run length matrices: short run emphasis, long run emphasis, gray level nonuniformity, run length nonuniformity, and run percentage [45]. Chu et al. [18] then extended these features defining two additional features: low gray level run emphasis and high gray level run emphasis. Table 2.2 provides a list of these features. In this table, n is the number

Short run emphasis	$\frac{1}{n}\sum_i^G\sum_j^R\frac{M(i,j)}{j^2}$
Long run emphasis	$\frac{1}{n}\sum_i^G\sum_j^R M\left(i,j\right)j^2$
Gray level nonuniformity	$\frac{1}{n}\sum_{i=1}^G\left(\sum_{j=1}^R M\left(i,j\right)\right)^2$
Run length nonuniformity	$\frac{1}{n}\sum_{j=1}^R\left(\sum_{i=1}^G M\left(i,j\right)\right)^2$
Run percentage	$\frac{n}{p}$
Low gray level run emphasis	$\frac{1}{n}\sum_{i=1}^G\sum_{j=1}^R\frac{M(i,j)}{i^2}$
High gray level run emphasis	$\frac{1}{n}\sum_{i=1}^G\sum_{j=1}^R M\left(i,j\right)i^2$

Table 2.2: Gray level run length features.

of rows and p is the number of pixels in the image. Furthermore, Dasarathy et al. [24] defined another four features on gray level run length matrices measuring the joint statistics of gray levels and run lengths.

Gray level run length method has also a considerable number of applications in automated diagnosis and grading of different types of cancer [57, 104, 103]. Doyle et al. [36] employed run length features together with co-occurrence features for the quantification of breast tissues. Albertgensen et al. [4] made use of these features for quantifying images of liver tissues. Moreover, Weyn et al. [103] and Bibbo et al. [10] used run length features for representing the chromatin texture of nuclei in colon epithelium for the classification of normal and cancerous colon tissues.

2.2.2.4 Multi Resolution Wavelets

Multi resolution wavelets has also a considerable amount of interest in the literature of texture analysis. Wavelet analysis is a mathematical method used for extracting information from data signals, including audio signals and images. Wavelets are functions that are used for representing data signals by dividing them into components of different scales [49, 25]. In this decomposition process, which is named wavelet transform, each resulting component of the analyzed signal is studied with a resolution that matches to its scale [26].

Wavelet transforms are divided into two groups which are Continuous Wavelet Transforms (CWT) and Discrete Wavelet Transforms (DWT) respectively. Use of wavelets in multiple scales allows achieving multi-scale representations which increases the effectiveness of the method. Features obtained from these multiwavelets such as short support, orthogonality, symmetry, and vanishing moments are known to be useful in image processing [3, 89, 95]. Wavelet analysis have a wide range of applications in literature, especially in the fields of image compression [7, 15, 23, 50, 61, 79] and image enhancement [73, 76]. Multiwavelet analysis method has also been employed in computerized diagnosis and grading of different types of cancers [91, 27, 101, 62, 104].

2.2.2.5 Image Fractals

Another textural method for image analysis is fractal geometry. Defined by Mandelbrot [71], a fractal is a geometrical shape which can be split into smaller parts, each of which looks identical to the whole. This method focuses on computing the fractal dimension of images, which gives an estimation of how a fractal appears to fill space as it is zoomed out.

Widely used method for computing fractal dimension is the box-counting method. According to this method, fractal dimension D, also known as Minkowski-Bouligand dimension [13], is defined in Equation 2.2 where $N(\epsilon)$ is the number of self-similar structures of diameter ϵ necessary to cover the structure.

$$D = \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log \frac{1}{\epsilon}} \tag{2.2}$$

Fractal geometry concept has been applied to analyze numerous textures in nature [75, 67]. It has also been involved in medical image analysis for the quantification of different types of tissues [106, 34, 16, 8, 39].

2.2.3 Structural Methods

Structural approach has been employed by a large number of studies in medical image analysis, in addition to its various applications in other fields of image processing. Structural methods characterize a tissue using the spatial distribution and neighborhood properties of its cellular components. To this end, representative graphs are constructed on these components, and a set of local and global features that provide quantitative information about tissues are computed.

Previous studies have proposed to use a number of different methods for constructing their graphs. Commonly used methods are Delaunay triangulations and their corresponding Voronoi diagrams, minimum spanning trees, Gabriel's graphs, and probabilistic graphs. These graph generation methods are summarized in the following subsections. Regardless of the methods, similar structural features have been extracted. A large number of these graph features are common for different types of graphs, and they are used by a majority of structural representations. Besides, features specific to particular graphs are also proposed. The common local and global graph features that are used by different structural methods are given in Table 2.3. In this table, it is also indicated which features can be used by which methods. Structural approaches employ these local and global graph features for the quantification of tissue images. First order statistics of local graph features such as mean and standard deviation are also used as global features of graph based representations.

2.2.3.1 Delaunay Triangulation

Delaunay triangulation is a commonly used graph generation method for modeling a wide range of concepts in different fields of science [41]. It was defined by Boris Delaunay [28], whose name was given to the method after his definition in 1934. Delaunay triangulation of a point set P in two dimensional space is defined as the

Feature	**Graph Category**
Area	VD, DT
Roundness	VD
Aspect ratio	VD
Circularity	VD
Number of sides	VD
Edge lengths	DT, MST, GG
Degree of nodes	DT, MST, GG, PG
Distance to the nearest neighbor	DT, MST, GG
Fractal dimension	MST
Size ratio of giant connected component to graph	PG
Eccentricity	DT, MST, GG, PG
Clustering coefficient	DT, MST, GG, PG
Shortest paths between nodes	DT, GG, PG
Diameter	DT, GG, PG, MST

Table 2.3: Structural features obtained from different graph based representations: Voronoi diagrams **(VD)**, Delaunay triangulations **(DT)**, Gabriel's graphs **(GG)**, minimum spanning tree **(MST)**, and probabilistic graphs **(PG)**.

set of triangles where each triangle conforms to the condition that there exist no point in the interior of the circle passing through the edge points p_x, p_y, $p_z \in P$ of the triangle. The condition of Delaunay guaranties that each triangle of the triangulation has an empty circumcircle and maximizes the minimum interior angles of the triangles. Figure 2.3 shows a sample Delaunay triangulation of 20 random points in two dimensional Euclidean space.

Delaunay triangulations have been used to build structural representations of various tissues for automated cancer diagnosis and grading. In [65], an automated framework that make use of Delaunay triangulations for grading of cervical intraepithelial neoplasia is proposed. Guillaud et al. [52] proposed to use Delaunay features for quantification of cervical tissues. Doyle et al. [36] proposed an automated diagnosis and grading method for breast cancer that also employs Delaunay triangulations.

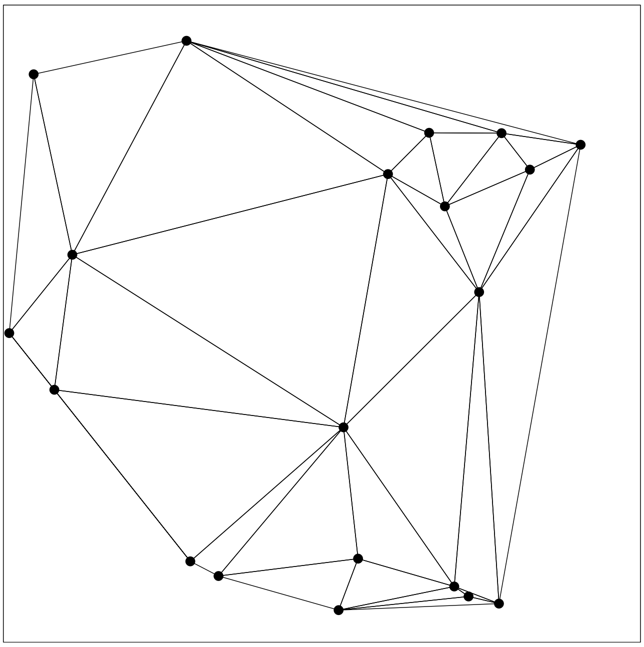


Figure 2.3: Delaunay triangulation of 20 random points in 2 dimensional Euclidean space.

2.2.3.2 Voronoi Diagrams

Voronoi diagrams, which are dual graphs of Delaunay triangulations, are also widely used in structural image analysis methods. The idea lying under Voronoi diagrams was first emphasized by Descarts [31, 74] when he was trying to model the solar system, claiming that the solar system consists of vortices, the convex regions around the stars. This concept has independently emerged in various fields including biology, physiology, chemistry, and physics. It has been used with different names in these field, such as *medial axis transform* in biology and physiology, *Wigner-Seitz zones* in chemistry and physics, *domains of action* in crystallography, and *Thiessen polygons* in meteorology and geography [78].

First formal definitions were proposed by the mathematicians Dirichtlet [32] and Voronoi [97, 98], with the names Dirichtlet tessellation and Voronoi diagram which have become the standard names in literature. Voronoi diagram is defined as the decomposition of a regular plane of n points into convex polygons called Voronoi regions such that each point inside a Voronoi region is closer to the

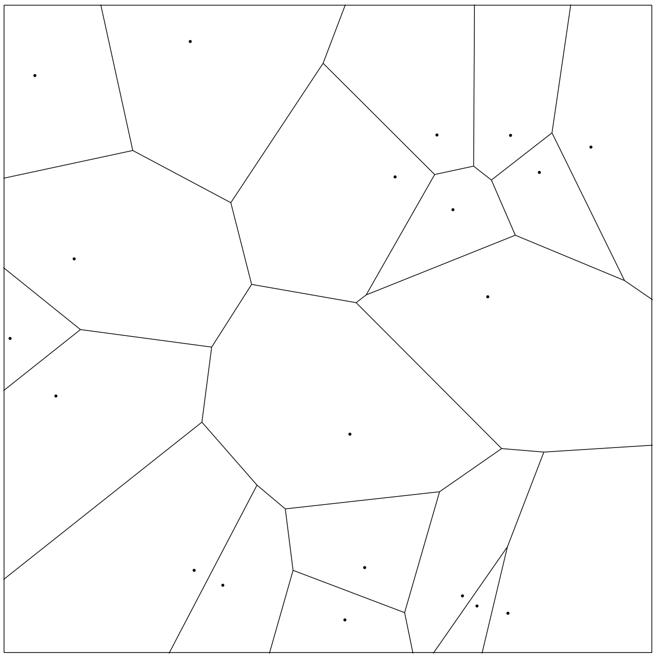


Figure 2.4: Voronoi diagram of the point set used in Figure 2.3.

origin point of that region than any other point in the plane. Voronoi diagrams are dual graphs of Delaunay triangulations, and they are generally used together in structural methods. Figure 2.4 shows the corresponding Voronoi diagram of the Delaunay triangulation given in Figure 2.3.

In literature, there are a number of studies that make use of Voronoi diagrams for quantification of histopathological images [90, 80]. Weyn et al. [102] proposed to use Voronoi diagrams to develop an automated method for diagnosis and grading of malignant mesothelioma. Doyle et al. [37] used Voronoi diagrams to extract structural image features from breast tissue images. Morover, a comparative study of morphological, textural, and structural image features that use Voronoi diagrams is provided in [104].

2.2.3.3 Gabriel's Graph

Another type of graph representations that is used for modeling the neighborhood characteristics of cellular components of tissues is Gabriel's graphs. This graph

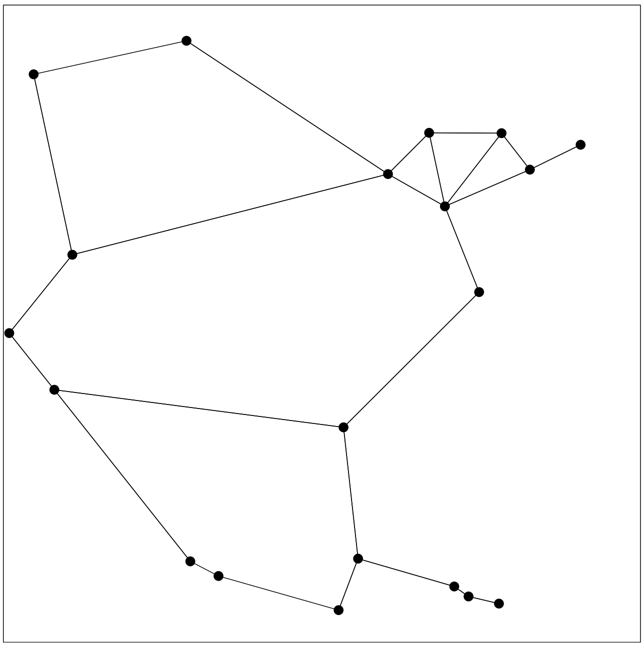


Figure 2.5: Gabriel's graph of the point set used in Figure 2.3.

generation method was proposed by Gabriel et al. [44]. Given a point set P, two points $p_i, p_j \in P$ are said to be Gabriel neighbors if and only if the circle having the line segment $[p_i p_j]$ as its diameter has no other point $p_k \in P$ in its interior. And the set of edges connecting the points that conform this property of Gabriel neighborhood is called a Gabriel's graph. A sample Gabriel's graph generated using point set of Figure 2.3 is presented in Figure 2.5. Gabriel's graphs are subgraphs of Delaunay triangulations [77]; this can be observed from Figure 2.3 and Figure 2.5.

2.2.3.4 Minimum Spanning Tree

Although it has less applications, minimum spanning trees are also employed for improving the graph based representations. A minimum spanning tree T of a weighted undirected graph G is the subgraph of G which has the minimum total weight among all possible connected subgraphs of G which have $n-1$ edges between n nodes of the graph without cycles and self loops. Figure 2.6 shows a minimum spanning tree of the Delaunay triangulation presented in Figure 2.3.

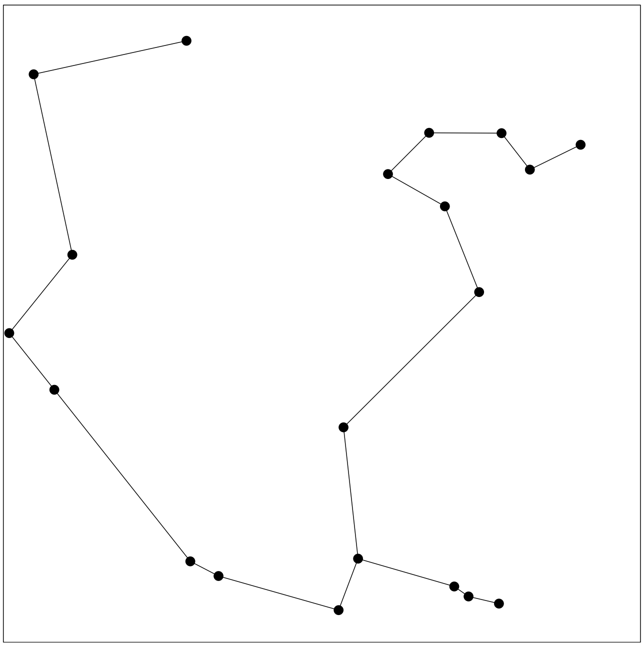


Figure 2.6: Minimum spanning tree of the Delaunay triangulation given in Figure 2.3.

Minimum spanning trees are used in conjunction with aforementioned graph based representations for automated cancer diagnosis and grading. Therefore, minimum spanning trees are obtained from Delaunay triangulations and a set of global features based on edge lengths are extracted. First order statistics such as mean, standard deviation, and kurtosis of these features are computed for quantifying tissues [104]. Weyn et al. [102] used minimum spanning trees in the automated diagnosis of malignant mesothelioma. Doyle et al. [36] and Guillaud et al. [52] also employed minimum spanning trees to quantify images of cervical tissues. Choi et al. [17] proposed to use minimum spanning trees in their comparative study on object-based, textural, and structural methods in the classification of bladder cancer.

2.2.3.5 Probabilistic Graphs

In probabilistic graphs edges between nodes are probabilistically assigned according to the distance between the nodes. In this method, the edge density of a graph is controlled with a probability function that can be defined in different ways. For

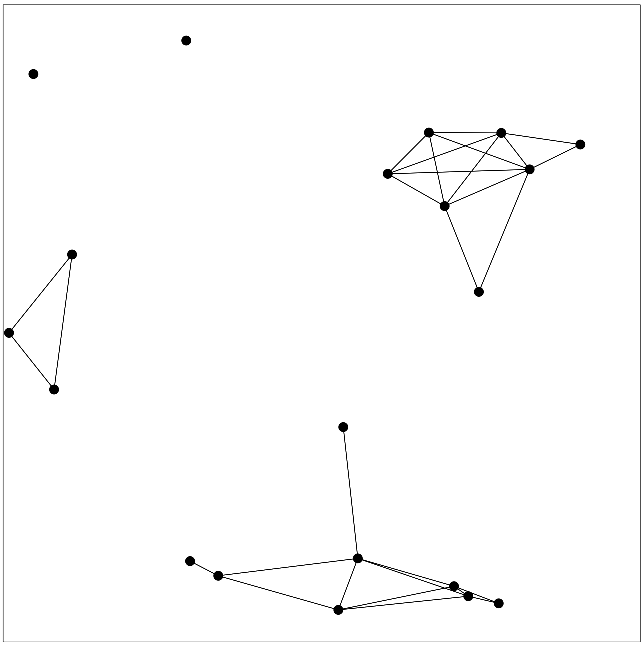


Figure 2.7: Probabilistic graph generated on the point set given in Figure 2.3. Here, Equation 2.3 is used with $\alpha = 2$ and $r = 0.2$.

example, in [30, 53], probabilistic graphs are constructed by using the following equation:

$$E\left(u,v\right) = \begin{cases} 1, \text{ if } d\left(u,v\right)^{-\alpha} > r \\ 0, \text{ otherwise} \end{cases} \tag{2.3}$$

In this equation, r is a random number between 0 and 1, and $d\left(u,v\right)$ is the Euclidean distance between nodes u and v. Here α is used as a parameter for controlling the edge density of generated graphs. Figure 2.7 shows a sample probabilistic graph which is generated using the function given in Equation 2.3.

Probabilistic graphs have been employed for modeling a large number of domains such as social networks and the world wide web. In [53], a novel probabilistic graph generation method is proposed to model brain tumor. Local and global features computed from these graphs are used for characterizing cancerous brain tissue [53, 30]. Demir et al. [29] proposed to use augmented cell-graphs, that are undirected, weighted, and complete probabilistic graphs without self loops, for automated diagnosis of cancer, where all possible edges between each pair of nodes are included in the graph, preventing the loss of any existing spatial information. In these graphs, edge weights are defined as the Euclidean distances between their end nodes. Furthermore, Gunduz-Demir [54] investigated the relation of different phases of cell-graphs with the malignancy of the cancer using graph evolution technique.

Chapter 3

Methodology

In Chapter 2 we have described the computational methods that are proposed for developing computer aided image analysis tools. These tools can be employed in cancer diagnosis and grading processes to obtain quantitative information about the characteristics of tissues, and help reducing the subjectivity of pathologists. Therefore, proposed methods define mathematical representations of tissues with the use of morphological, textural, and structural features of histopathological images. Although these methods yield accurate results for diagnosis and grading of cancer, they mainly focus on nuclear characteristics of tissues, ignoring the role of other components in the tissue structure. In this thesis, we propose a novel structural representation for the quantification of different components of histopathological tissue images. In addition to nuclear tissue components, the proposed method considers different components of a tissue such as luminal and stromal regions. It takes spatial relationships and neighborhood properties of these components into account in making decisions.

The proposed method consists of the following four steps: (1) node identification, (2) edge identification, (3) feature extraction, and (4) classification. In the node identification step, image pixels are converted into Lab color space and then segmented into three sets corresponding to luminal, stromal, and nuclear regions. Subsequently, morphological operations are applied on each of these pixel sets for reducing noise and circular components are located on each region. These

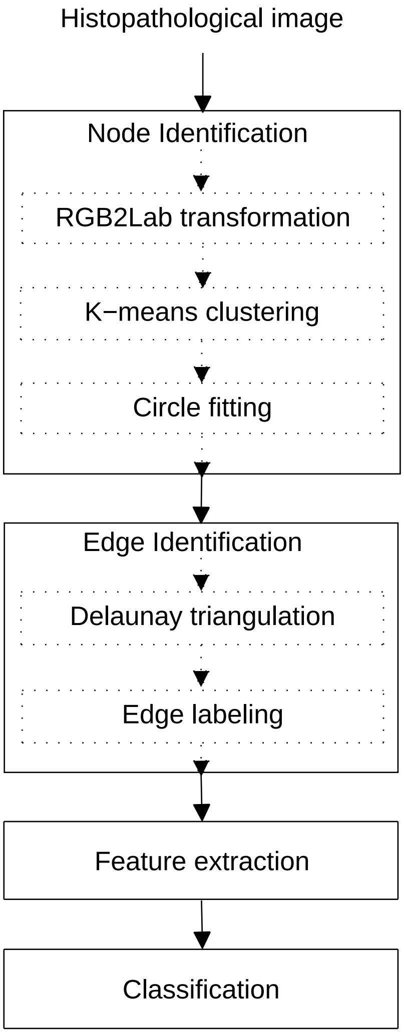


Figure 3.1: Overview of the color graph approach.

components are considered as the graph nodes. In the next step, a graph is constructed on these nodes making use of Delaunay triangulation, and the triangle edges are colored according to the component types of their end nodes. Then, a set of graph features is computed from the constructed color graph that is used in the classification step. Figure 3.1 shows a shematic representation of the proposed approach. Next subsections provide detailed descriptions and outputs of the steps of the proposed color graph approach.

3.1 Node Identification

Color graph approach takes all components of a tissue into account to effectively quantize its histopathological image. To this end, nuclear, stromal, and luminal components of a tissue are mapped to graph nodes. To perform this mapping, histopathological images should be segmented into its cytological components. The ideal way of this segmentation is to identify the exact boundaries of each component. However, in a typical histopathological image, there could be staining and sectioning related problems, including the existence of touching and overlapping components, lack of dark separation lines between a component and its surroundings, inhomogeneity of the interior of a component, and presence of stain artifacts in a tissue [48]. This complex nature of a histopathological image scene leads to a difficult segmentation problem even for the human eye. Therefore, the proposed approach approximately represents the tissue components with three sets of circular primitives corresponding to nuclear, stromal, and luminal regions. The centroids of these tissue components are considered as the node locations in constructing a color graph.

3.1.1 Clustering with K-means

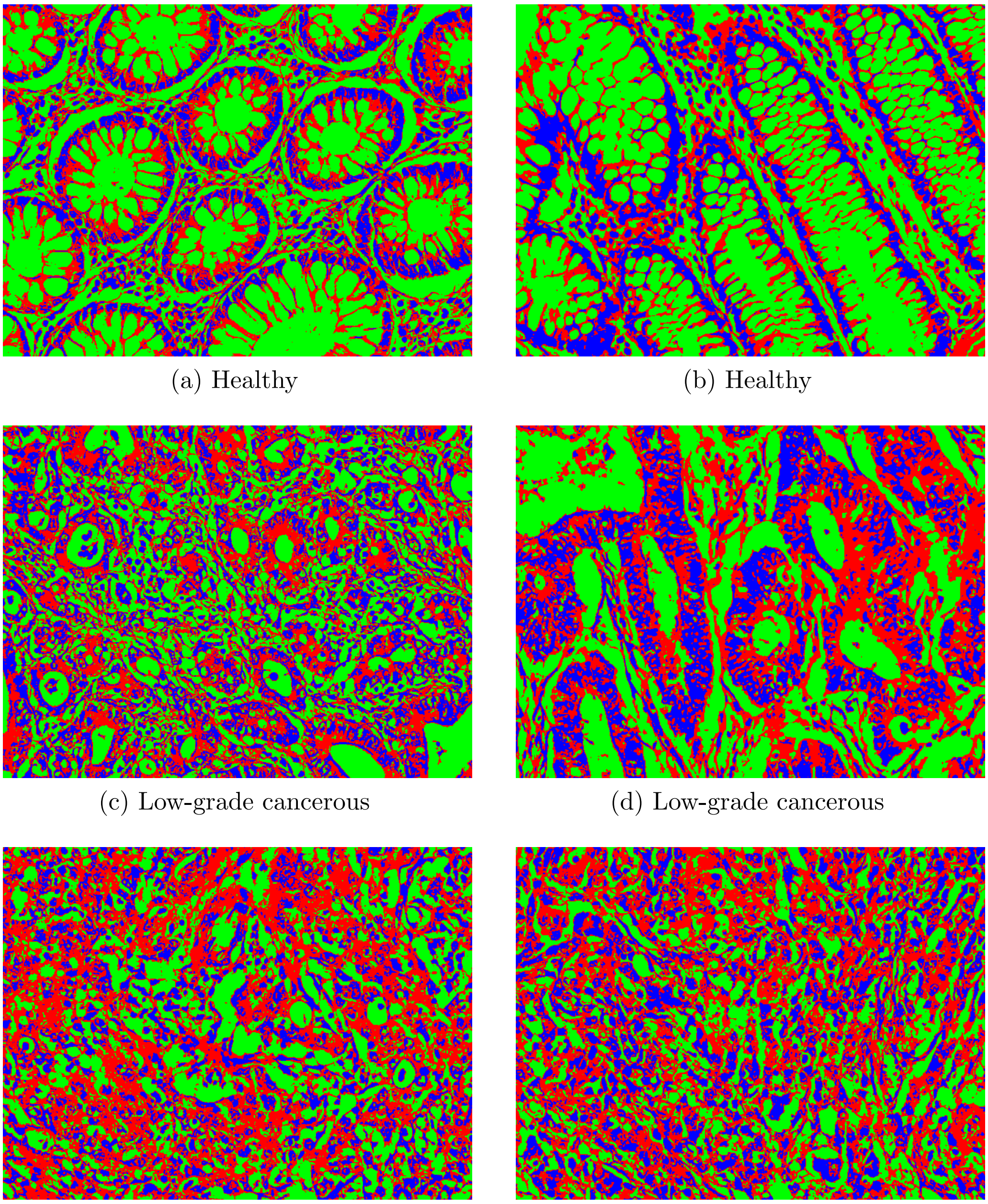
To identify the component of a tissue, pixels of its histopathological image are segmented into three clusters using the k-means algorithm. The number of clusters is particularly selected as three since there are mainly three color groups in the histopathological image of a tissue that is stained with hematoxylin-and-eosin. These colors are purple, pink, and white, which typically correspond to nuclear, stromal, and luminal components, respectively.

Before clustering, image pixels are converted into Lab color space to take advantage of luminance components of pixel values. Lab color space is a commonly preferred color space over RGB color space in image processing applications. It is a color-opponent space, which approximates the visual perception of human eye. Its dimensions L, a, and b correspond to luminance, greenness-redness, and blueness-yellowness components respectively. This color space improves image analysis methods since its L component closely matches human perception of lightness [47].

The k-means algorithm [70], which is one of the most commonly used unsupervised learning algorithms that partitions an N dimensional data into k subsets, is used for segmentation. This algorithm attempts to assign each data point to a cluster S_i, minimizing the *within cluster sum of squares* or *squared error function* E given in Equation 3.1.

$$E = \sum_{i=1}^{k} \sum_{j \in S_i} \parallel x_j - c_i \parallel^2 \tag{3.1}$$

In Equation 3.1, the term $\parallel x_j - c_i \parallel$ is a distance measure between the centroid c_i and the observation x_j assigned to the cluster S_i. In this study, we use k-means implementation of Matlab with Euclidean distance as the distance measure for clustering. With the utilization of k-means algorithm, the centroids of clusters are computed and types of these centroids are determined according to the values of their L component. This determination depends on the idea that lower values of L component correspond to darker colors, and higher values correspond to brighter colors. Thus, we label the pixels that have the lightest centroid as *luminal* pixels since luminal regions of a tissue correspond to white colored pixels in the histopathological image. Similarly, pixels that have the darkest centroid are labeled as *nuclear* pixels since nuclear tissue components are the darkest colored regions in images. And the remaining pixels are labeled as



(a) Healthy

(b) Healthy

(c) Low-grade cancerous

(d) Low-grade cancerous

(e) High-grade cancerous

(f) High-grade cancerous

Figure 3.2: Pixel clusters of histopathological images given in Figure 2.2 that are obtained using k-means.

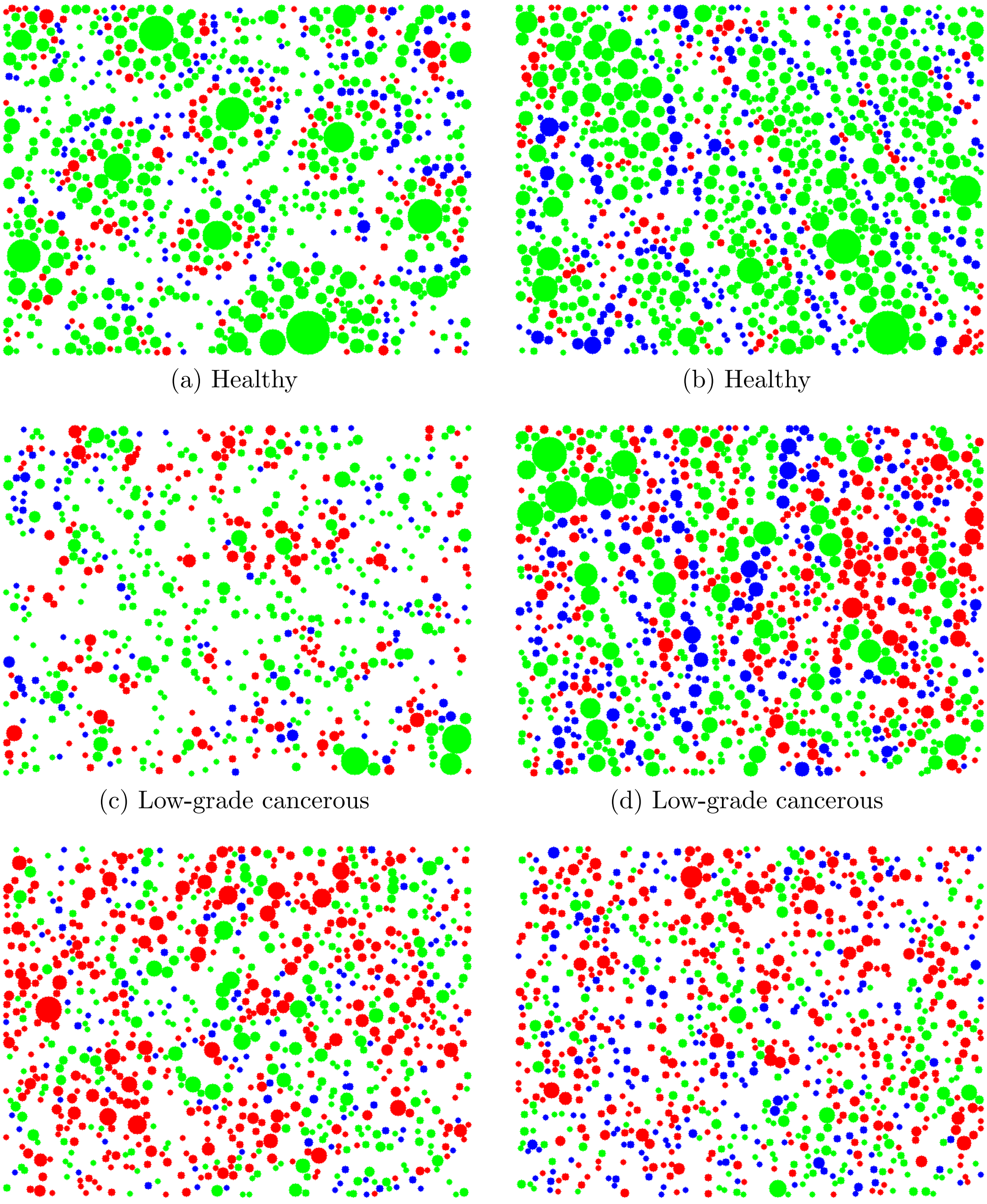
stromal pixels. Figure 3.2 shows resulting images of the clustering process.

At the end of this step, we obtain three binary images corresponding to nuclear, stromal, and luminal regions, respectively. These images are preprocessed to reduce the noise that arises from the incorrect clustering of pixels. To this end, on the pixels of these images, closing and opening operators with a square structuring element of size 3 are applied sequentially. Closing operation removes noisy pixels and fills small holes, whereas opening removes small objects from the images. After this preprocessing, these binary images are used as pixel masks on nuclear, luminal, and stromal regions.

3.1.2 Circle Fitting

After clustering and preprocessing of histopathological images, we obtained three images each of which consists of pixels that belong to nuclear, luminal, and stromal regions. This segmentation process discriminates different types of tissue components. However, it still does not eliminate the difficulty of identifying the exact boundaries of individual tissue components. Thus, segmentation of these components remains as a challenging problem.

In order to overcome this difficulty, the proposed approach approximately represents a tissue component with a circle object instead of identifying its exact boundary. To perform this task, it employs a heuristic method called *circle fitting* [94, 55] for defining circular primitives on binary images obtained in the previous step. Given a particular connected component of the image, the algorithm first locates a circle on the pixels of this component, which has the largest diameter conforming that it does not exceed the borders of the component. The pixels covered by the located circle are labeled. The algorithm then locates smaller circles on the unlabeled pixels of the connected component conforming that the circles inside the component do not overlap each other and do not exceed the borders of the component. The algorithm iteratively locates smaller circles on unlabeled pixels as far as the areas of generated circles reach to a value smaller than a predefined area threshold. This iterative procedure continues until each



(a) Healthy

(b) Healthy

(c) Low-grade cancerous

(d) Low-grade cancerous

(e) High-grade cancerous

(f) High-grade cancerous

Figure 3.3: Circular primitives that are obtained by applying the circle fitting algorithm to pixel clusters of images given in Figure 3.2.

connected component of the image is processed.

Using the aforementioned method, pixel masks of nuclear, stromal, and luminal regions are transformed into circular primitives that approximately represent these tissue components. The motivation lying under the idea of using circular objects is that cytological components of the colon tissue components typically have curvy boundaries and circles are efficiently located on a set of pixels. This approach effectively maps a tissue component to a circular primitive; however, it may also split a tissue component into more than one circle especially those having elliptical shapes. Figure 3.3 shows the circles that are obtained by transforming the images given in Figure 3.2 into circular primitives with the use of the circle fitting algorithm.

At the end of circle fitting step, we obtain three sets of circles corresponding to different types of tissue components. The coordinates of the centroids and types of these circles are used the next step of the proposed approach for defining the edges of the graph.

3.2 Edge Identification

In the previous section, we have identified the locations of circular primitives as graph nodes. The next step of the proposed approach assigns edges between these nodes to build a graph model that represents the underlying tissue image. For this purpose, it constructs a Delaunay triangulation on these nodes.

As mentioned in Section 2.2.3, Delaunay triangulation method is a commonly used structural method for automated analysis of histopathological images. The definition of Delaunay triangulation models the spatial relationships of nodes in a graph. This characteristic of Delaunay triangulation makes it suitable for representing organizational structures of tissues.

Previous studies on structural tissue analysis commonly proposed to construct Delaunay triangulations only on the nuclear tissue components focusing on the

Green	:	Luminal-luminal edge
Red	:	Stromal-stromal edge
Blue	:	Nuclear-nuclear edge
Yellow	:	Luminal-stromal edge
Cyan	:	Luminal-nuclear edge
Magenta	:	Stromal-nuclear edge

Table 3.1: List of colors that are used for labeling triangle edges.

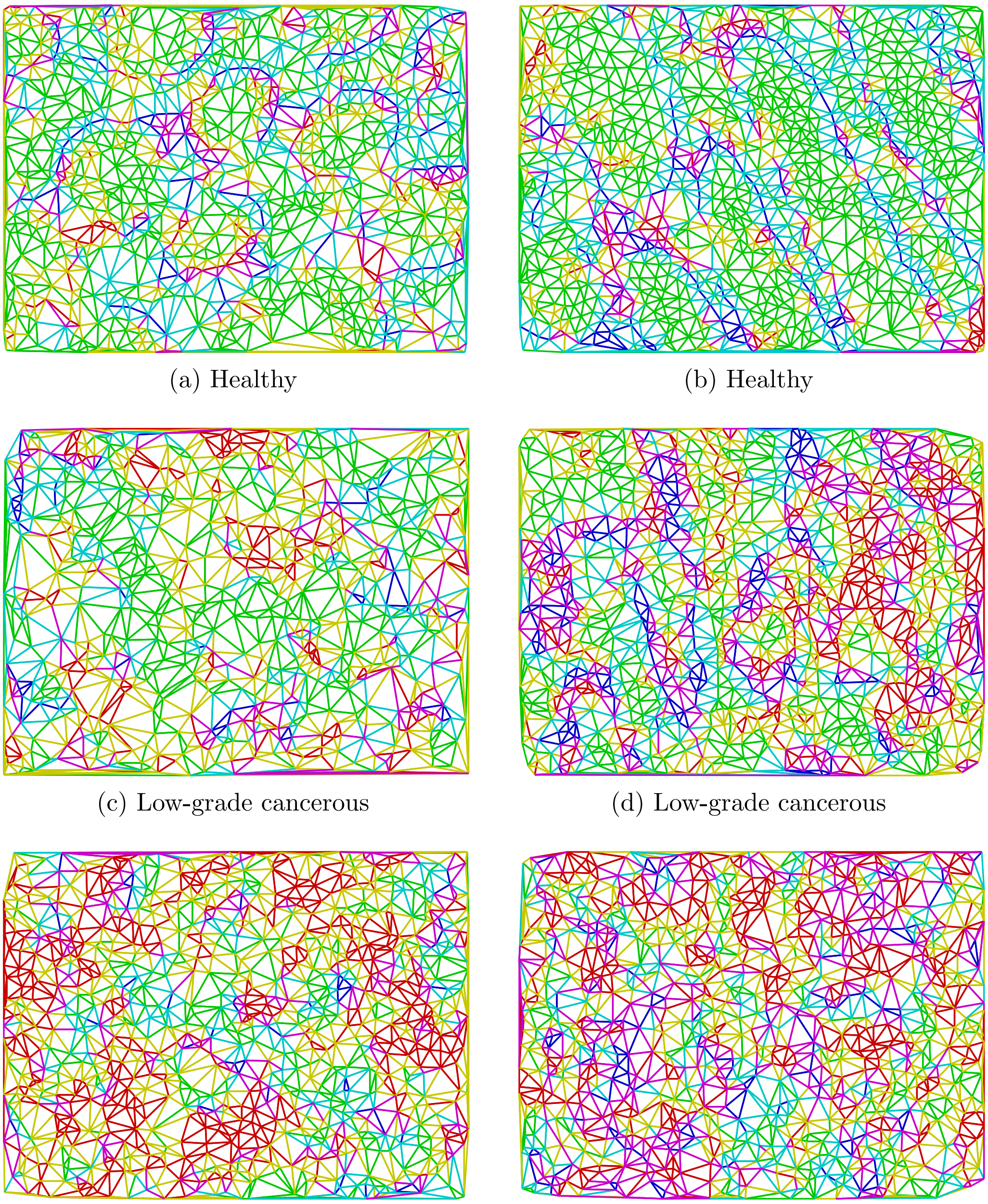
distribution of cell nuclei in tissues. Although these mathematical representations provide representative information for some tissue types, they may not efficiently represent tissues where cytological components other than nuclei also reflect the structural characteristics of the tissue.

On the other hand, the proposed approach builds a Delaunay triangulation on three types of nodes corresponding to nuclear, stromal, and luminal tissue components. Then, it labels the edges of the resulting graph according to the types of their end nodes. As there are three different types of nodes, edges are colored with one of the six colors that are given in Table 3.1. Figure 3.4 shows the color graphs that are constructed on the circles given in Figure 3.3.

In literature, there have been a number of methods for constructing Delaunay triangulations [42, 43, 51]. In this study, we have employed a Matlab function that uses the Qhull [9] algorithm for computing Delaunay triangulation.

3.3 Feature Extraction

In order to quantify histopathological images, three groups of global graph features that are commonly used by structural image analysis methods are computed from constructed color graphs. These are average degree, average clustering coefficient, and diameter. Additionally, we aim to introduce the color information into the computations of these graph features. The definitions of these features are given in the following subsections.



(a) Healthy

(b) Healthy

(c) Low-grade cancerous

(d) Low-grade cancerous

(e) High-grade cancerous

(f) High-grade cancerous

Figure 3.4: Color graphs that are generated from the graph nodes given in Figure 3.3 and that are colored with one of the six colors given in Table 3.1 depending on the component types of their end nodes.

3.3.1 Average Degree

In graph theory, the degree of a node is defined as the number of its adjacent neighbors. The average degree of a graph is the mean of the degrees computed for every node in the graph. In this study, seven degrees are computed for a single node using the following definitions:

- Degree: Given a color graph $G\left(N, E\right)$, the degree d_i of node $i \in N$ is the total number of its edges, regardless of their colors. This is a common feature that is used to quantify the local connectivity properties of a graph.
- Color degree: Given a color graph $G\left(N, E\right)$, the color degree ${d_i}^c$ of node $i \in N$ is the total number of its edges that are of color c. For a particular node of the color graph, we define six color degrees as there exits six types of edges in the graph connecting three types of nodes. This feature is used for quantifying the local connectivity of different types of graph nodes.

Taking the averages of these local graph features, seven average degrees are obtained for an individual color graph.

3.3.2 Average Clustering Coefficient

Average clustering coefficient also provides information about the connectivity of a graph [35]. It indicates the density of connections in the neighborhood of a node. For an individual color graph, four clustering coefficients are computed using the following definitions:

- Clustering coefficient: Clustering coefficient of a node is defined as the ratio of the number of existing edges over the number of all possible edges between its neighbors. Clustering coefficient C_i of node i is computed as follows:

$$C_i = \frac{2d_i \cdot E_i}{d_i \cdot (d_i - 1)} \tag{3.2}$$

where d_i is the degree of i and E_i is the number of existing edges between its neighbors.

- Color clustering coefficient: The color clustering coefficient of a node is defined to quantify the clustering information of nodes of the same component type. For this purpose, a colored clustering coefficient is defined for each node considering only its neighbors of the same type. For instance, for a luminal component node, only the luminal-luminal edges are considered and a luminal clustering coefficient ${C_i}^l$ is defined as

$$ {C_i}^l = \frac{2{d_i}^l \cdot {E_i}^l}{{d_i}^l \cdot \left({d_i}^l - 1\right)} \tag{3.3} $$

  where ${d_i}^l$ is the number of luminal neighbors of i and ${E_i}^l$ is the number of existing edges between these neighbors. Similarly, color clustering coefficients ${C_i}^s$ and ${C_i}^n$ are computed for stromal and nuclear component nodes.

Averaging the clustering coefficients over all nodes, four global features are obtained to quantify a color graph.

3.3.3 Diameter

The shortest path between two nodes u and v of a graph is defined as the minimum number of adjacent edges that have to be traveled from u to v. The diameter of a graph is the length of the longest of the shortest paths between any pair of graph nodes. Diameter is a global feature that indicates the size of the graph. In this work, seven diameters, one colorless and six color diameters, for a color graph are computed. The first diameter is computed without considering the edge colors such that adjacent edges in the computed paths may be of different types. The other six of the graph diameters are computed considering only the edges with a color of interest. For instance, the green diameter is the longest of the shortest paths between any pair of nodes that are reachable to each other using only the luminal-luminal (green) edges.

3.4 Classification

In Section 3.3, 18 dimensional feature vectors are defined for color graphs to quantify histopathological tissue images. In the last step of the proposed approach, these feature vectors are used by support vector machines to achieve classification.

3.4.1 Support Vector Machines

For an effective classification method, it is important to accurately classify not only the observed data but also the unknown data. Thus, it is necessary to select the most appropriate boundary that will optimally separate the unpredicted data. Numerous parametric models that are based on the probability density estimations of classes are proposed for performing this task. Support vector machines provide a non-parametric classification method that solves an optimization problem.

The support vector machine is a kernel-based supervised learning method that is used for classification and regression. The support vector machine algorithm was originally proposed as a linear classifier [21]. The algorithm aims to partition data points of two classes in n dimensional space with an $n-1$ dimensional hyperplane. In general, there may be more than one hyperplane separating the data points. Figure 3.5a shows such possible separations of data points of two classes in two dimensional space. The fundamental idea of the algorithm is to find the separating hyperplane that maximizes the margin, which is the distance of the nearest data points in both data sets to the separating hyperplane. Figure 3.5b shows an optimal separation of data points in Figure 3.5a, which maximizes the margins. Data points that lie on the margins are called support vectors.

The original definition of support vector machines provides a linear classification model. However, it is not always possible to linearly separate data. For example, for the data points given in Figure 3.6a, it is impossible to locate a linear boundary that totally separates two classes, which limits the effectiveness of support vector machines. On the other hand, extensions have been proposed to

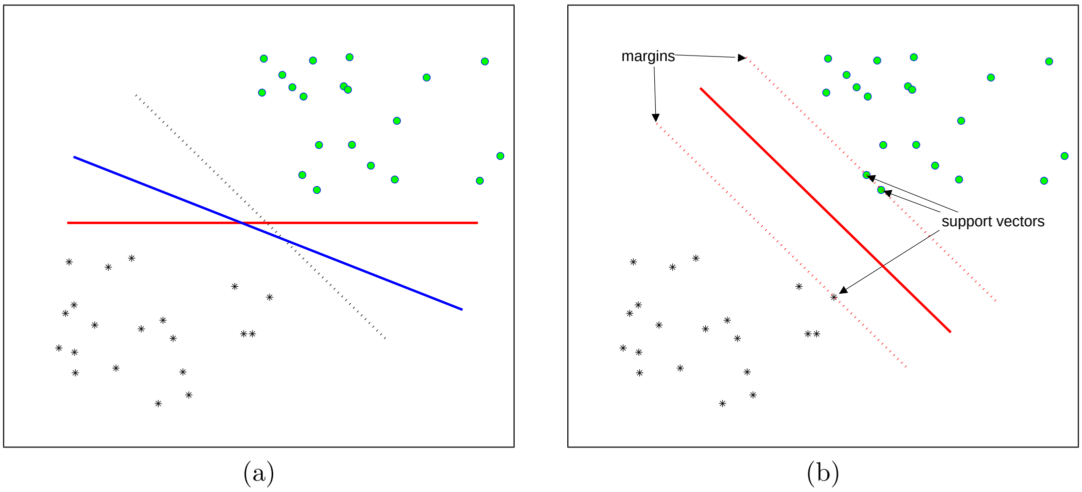


Figure 3.5: (a) Three possible separating lines, and (b) the separating line with maximum margin.

overcome this limitation of support vector machines [12]. For this reason, kernel method is used for mapping data points to a higher dimensional space, in which the data set is linearly separable. In this method, data points in the original feature space are transformed to feature points in a higher dimensional space with the use of a kernel function. Figure 3.6b demonstrates the linear separation boundary in the transformed feature space. Common kernel functions used by support vector machines are linear, Gaussian radial basis function, polynomial, and sigmoid kernels.

Another limitation of support vector machines is that they are designed to solve binary classification problems and they are not suitable for multi-class classification. On the other hand, methods have been proposed to utilize support vector machines for multi-class classification problems [60]. These efforts are mainly based on reducing a multi-class problem to multiple binary classification problems [22]. One-to-one [66] and one-against-all [96] classification schemes are examples of these efforts. Furthermore, there have been studies that propose to generalize support vector machines for multi-class problems [100, 96, 22].

For the classification of histopathological images, we use an implementation of support vector machines provided at `http://www.csie.ntu.edu.tw/`

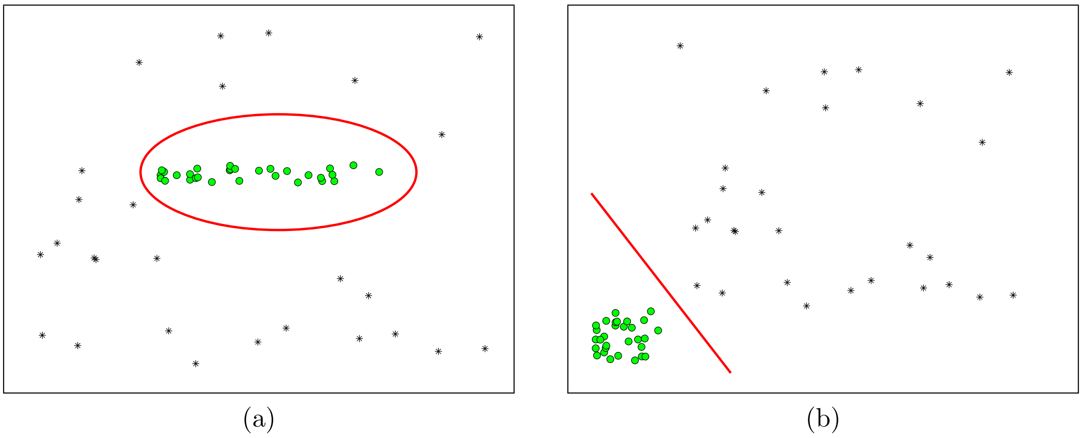


Figure 3.6: (a) Data points in two dimensional space that are non-linearly separable, and (b) the data points obtained by transforming these points using kernel method.

`~cjlin/bsvm/index.html`; this is an implementation of algorithms discussed in [60]. By default, the classifier uses a Gaussian radial basis function kernel with Gamma value $1/k$, where k is the dimension of the feature space. We set all of its learning parameters to their default values except the value of parameter C, which is the cost parameter that affects the trade-off between the training error and the margin. In the experiments, the value of C is selected using 10-fold cross-validation on the training set. In particular, a candidate set of $\{1, 2, ..., 9, 10, 20, ...90, 100, 150, ..., 950, 1000\}$ is considered for C and the one that leads the best cross-validation accuracy is selected.

3.4.2 Feature Selection

In the training stage of a classification process, it is sometimes possible for the classifier to summarize the observed data, which is named as the over-fitting problem. This problem may occur when the number of features is relatively high with respect to the size of training data. In the proposed approach, 18 color graph features are defined and used for the classification of histopathological images. This large number of features may result in the over-fitting problem. Therefore, we make use of two feature selection algorithms for investigating the effectiveness of the color graph features, that are forward selection and backward elimination. The details of this process is described below.

3.4.2.1 Sequential Forward Selection

Sequential forward selection method is an iterative greedy search algorithm used for selecting a subset of features [56, 69]. The algorithm starts with an empty set F_0 of selected features and a set C of candidate features. In each iteration, it includes the most discriminative feature of the candidate feature set to the set of previously selected features. To this end, at each level l_i, it finds the feature $f \in C$ that maximizes the classification accuracy when included in the feature set F_{i-1} of the previous iteration. In each iteration, 10-fold cross validation on the training set is used for measuring the classification accuracy. The selected feature is then removed from the candidate set C and included into the set F_i of selected features.

3.4.2.2 Backward Elimination

Backward elimination is another greedy search algorithm, which is used for feature selection [56]. It also proposes an iterative approach similar to forward selection. However, backward elimination algorithm iteratively eliminates insufficient features from an initial set, as opposite to forward selection. It excludes the least descriptive feature from this initial set in each level by finding the candidate feature that results in the minimum classification accuracy decrease when excluded from the feature set. This process eliminates the least effective feature from the initial feature set at each iteration. Removal of features increases the accuracy for a number of iterations, and then the accuracy strictly decreases when a minimal set of most discriminative features is reached. On the other hand, exclusion of a feature may not always increase the classification accuracy.

Chapter 4

Experiments and Results

This chapter describes the experimental methodology and provides the results of the proposed approach for a three class classification problem of colon cancer. Analytical experiments on the effectiveness of the color graph approach and its comparisons with other structural methods are also given.

4.1 Dataset

The data set used in this study consists of 3236 histopathological images of colon tissues taken from 258 different patients. These patients are randomly selected from the Pathology Department archives in Hacettepe University School of Medicine. Each tissue section in the data set has a thickness of 5 microns. These samples are stained with the hematoxylin-and-eosin technique. Each histopathological image in the data set is labeled as normal, low-grade adenocarcinomatous, or high-grade adenocarcinomatous colon tissues by a medical expert.

The data set is randomly divided into training and test sets. The training set consist of 1644 images and includes 510 normal, 859 low-grade adenocarcinomatous, and 275 high-grade adenocarcinomatous tissue samples taken from 129 patients. The test set consist of 1592 images and includes 491 normal, 844 low-grade adenocarcinomatous, and 257 high-grade adenocarcinomatous tissue samples taken from the remaining 129 patients. The training set is also divided into 10 folds to be used in cross validation. Here, every sample of a particular patient is included into the same fold to avoid interdependent subsets. Moreover, the numbers of samples for each class in the training set is relatively unbalanced, such that, the number of low-grade samples is very large with respect to the numbers of samples of the other classes. Nevertheless, this unbalanced distribution of samples over classes may introduce a drawback in the learning process. In order to overcome this drawback, an appropriate number of samples in the training set are duplicated to compensate the sample numbers for each class.

The tissue images that are used in the experiments are taken using a Nikon Coolscope Digital Microscope with 20× objective lens. This magnification is high enough to obtain homogeneous images and at the same time low enough to obtain images containing multiple glands. The image resolution affects the computational time required for processing a single image. Thus, the image resolution is selected as 480×640, which gives both accurate classification results and relatively lower computational times.

4.2 Classification Results

In this study, we have focused on the discrimination of normal, low-grade cancerous, and high-grade cancerous tissues. This section gives the results for the classification of these types of colon tissues using support vector machines (SVM). First, confusion matrices are given for the classification of training and test data with the use of 18 features computed from color graphs. Then, classification results for features that are selected using forward selection and backward elimination are provided.

		Computed Class			Accuracy (%)
		Normal	High-grade	Low-grade	
Actual Class	Normal	**503**	3	4	98.63
	Low-grade	8	**797**	6	92.78
	High-grade	0	15	**260**	94.55
				Overall	**94.89**

(a)

		Computed Class			Accuracy (%)
		Normal	High-grade	Low-grade	
Actual Class	Normal	**462**	20	9	94.09
	Low-grade	14	**745**	85	88.27
	High-grade	8	45	**204**	79.38
				Overall	**88.63**

(b)

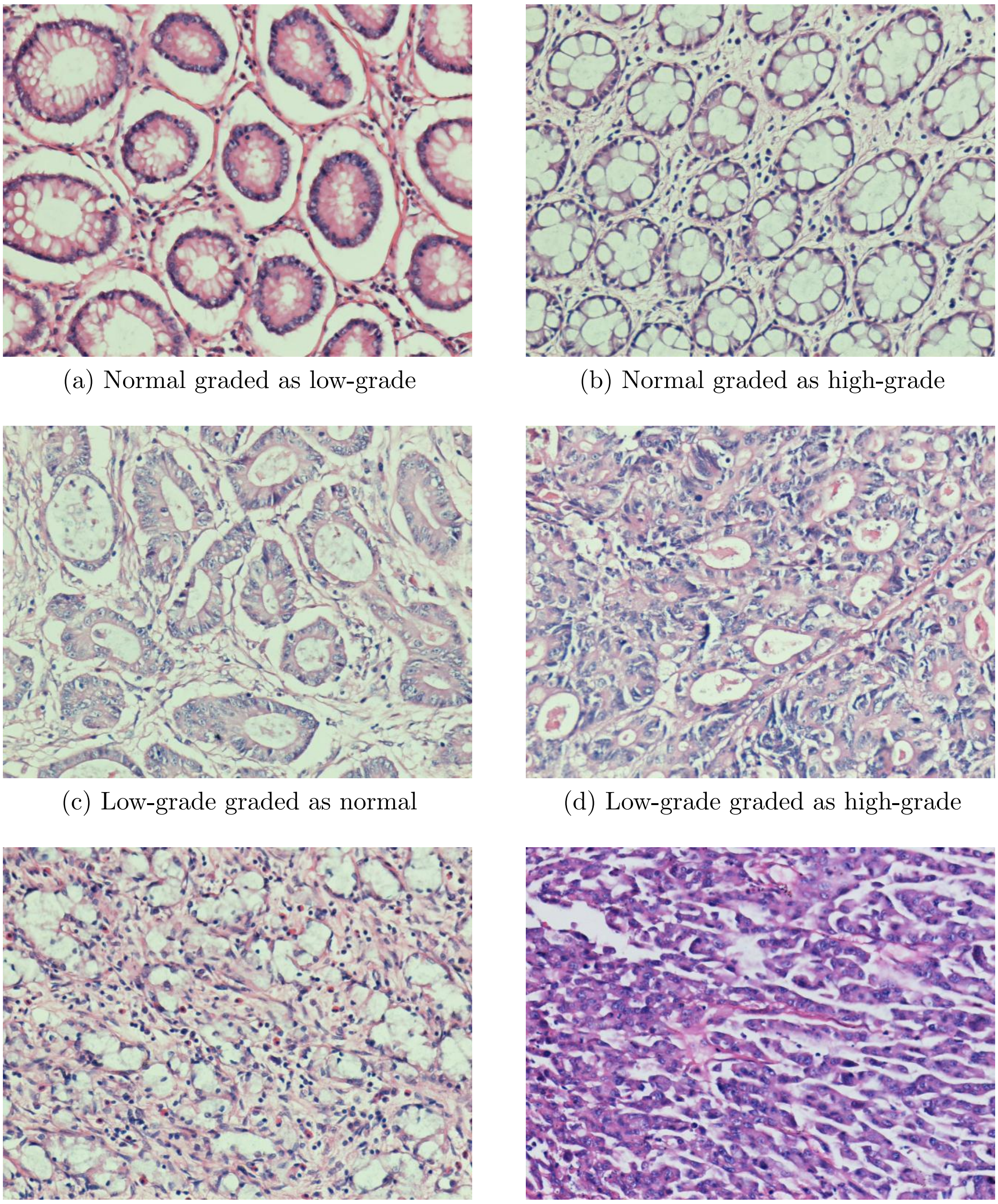
	Average	Std. dev.
Overall	88.63	0.14
Normal	94.09	0.16
Low-grade	88.27	0.17
High-grade	79.37	0.59

(c)

Table 4.1: Confusion matrices obtained with the color graph approach for (a) the training set and (b) the test set, and (c) average accuracy results and their standard deviations obtained by applying leave-one-patient-out cross validation on the test set. These results are obtained when all color graph features are used.

4.2.1 Results for Color Graph Features

In this section we provide classification results for the color graph approach. In the training of SVM, the cost parameter C is selected as 1 using 10-fold cross validation. The confusion matrices for the color graph approach are given in Table 4.1. As indicated in the table, the color graph approach yields high accuracies for both the training and test sets. The accuracies for all classes in the training set are also accurate. It is also shown that although the proposed approach leads to higher accuracies for the classification of normal and low-grade cancerous tissues in the test set, it gives relatively lower accuracies for the high-grade cancerous tissues.



(a) Normal graded as low-grade

(b) Normal graded as high-grade

(c) Low-grade graded as normal

(d) Low-grade graded as high-grade

(e) High-grade graded as normal

(f) High-grade graded as low-grade

Figure 4.1: Sample images of (a) - (b) normal, (c) - (d) low-grade cancerous, and (e) - (f) high-grade cancerous tissues that are misclassified by color graph approach.

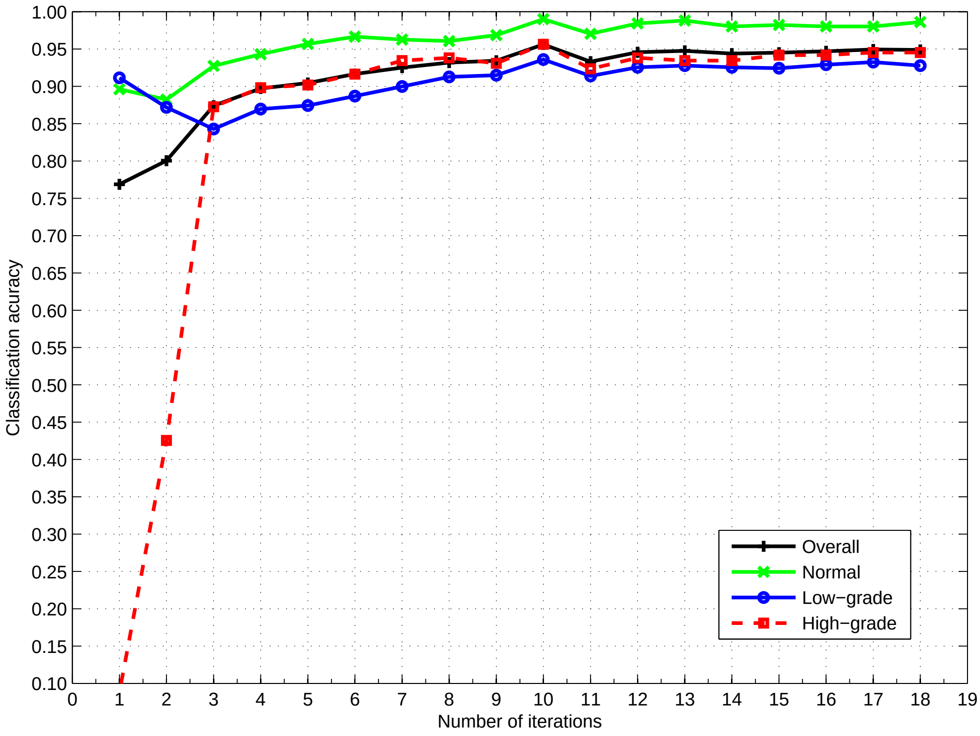
For this reason, we check the misclassified tissues of each class in the test set. Here, we observe that almost all of the normal tissues that are misclassified as low-grade comprise sectioning and staining related problems. Additionally, almost all of the normal tissues that are misclassified as high-grade consist of relatively large number of glands. Figure 4.1a and Figure 4.1b show the examples of normal tissues that are classified as low-grade and high-grade. We also observe that almost half of the misclassified low-grade tissues are the cases in which glands are well differentiated and the cases lying at the boundary between low-grade and high-grade cancer. Examples of misclassified low-grade tissues are given in Figure 4.1c and Figure 4.1d. On the other hand, we do not observe any common characteristics for the misclassified high-grade samples. Figure 4.1e and Figure 4.1f show the examples of such samples.

We also investigate how sensitive the proposed approach is to the data of a particular patient. For this purpose, we employ leave-one-patient-out cross validation on the test set: we exclude the samples of a particular patient from the test set and classify the remaining samples. This process is repeated for samples belonging to each of the 129 test patients. The averages and standard deviations of classification accuracies are given in Table 4.1c. This table shows that the results are stable and do not show too much differences with the inclusion/exclusion of a particular patient.

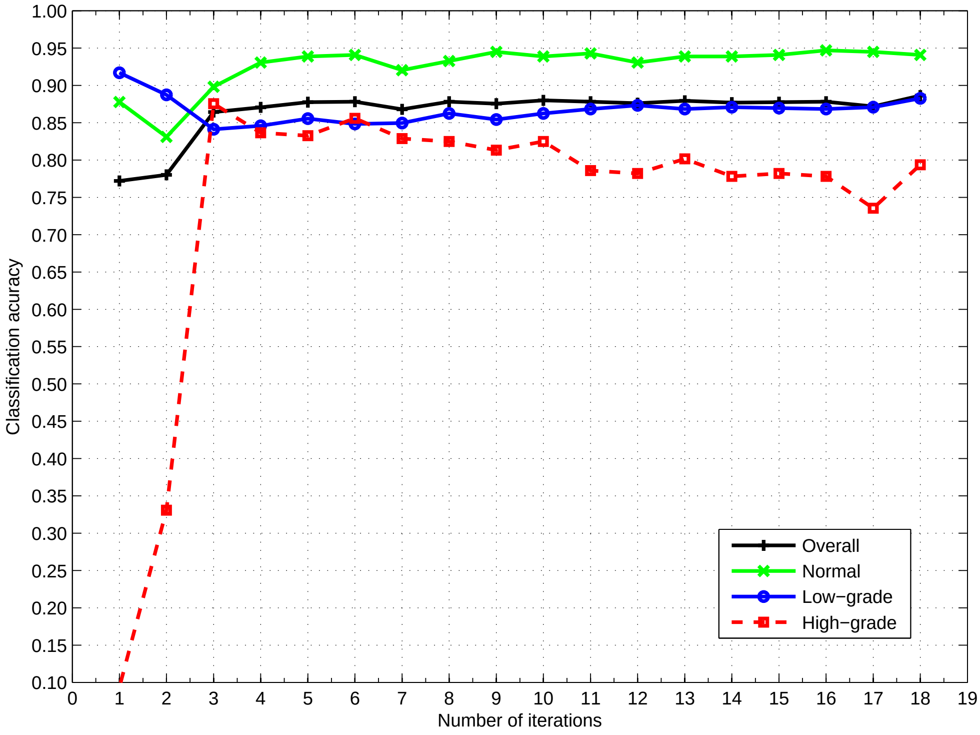
4.2.2 Results for Forward Selection

For the analysis of color graph features, we first employ sequential forward selection algorithm. In each iteration of the algorithm, a particular feature is selected and included in the classification for the following iterations. Figure 4.2 shows how the classification accuracies for the training and test sets change with respect to arbitrary iterations of sequential forward selection.

It is observed from Figure 4.2a that classification accuracies for all classes in the training set increase until iteration 10. For the following iterations, included features do not significantly improve the performance. In Figure 4.2b, it



(a)



(b)

Figure 4.2: Classification accuracies of (a) the training set and (b) the test set with respect to iterations of forward selection.

$Degree(red)$	Average degree for stromal-stromal edges
$Degree(green)$	Average degree for luminal-luminal edges
$Degree(blue)$	Average degree for nuclear-nuclear edges
$Degree(magenta)$	Average degree for stromal-nuclear edges
CC	Average clustering coefficient
$CC(luminal)$	Average clustering coefficient for luminal component
$CC(stromal)$	Average clustering coefficient for stromal component
$CC(nuclear)$	Average clustering coefficient for nuclear component
$Diameter$	Diameter
$Diameter(blue)$	Diameter for nuclear-nuclear edges

Table 4.2: List of features selected by forward selection.

is observed that at the beginning, there are improvements in the classification accuracies for all classes in the test set. However, after iteration 6, the accuracy for high-grade decreases as new features are included in classification, while accuracies of other classes do not significantly change.

Considering the results obtained on the training set, we select the iteration number as 10, where overall classification accuracies for both training and test sets are maximized. The features in this subset are given in Table 4.2. Confusion matrices and leave-one-patient-out classification results that are obtained using these features are also provided in Table 4.3. Here, C parameter of the support vector machine is selected as 20 using 10-fold cross validation.

4.2.3 Results for Backward Elimination

To investigate the effectiveness of color graph features, we also employ backward elimination algorithm. In each iteration, a particular feature is selected and excluded from classification for the following iterations. Figure 4.3 shows how the classification accuracies for training and test sets change with respect to arbitrary iterations of backward elimination.

Figure 4.3a shows that classification accuracies for all classes in the training set do not significantly change for a number of iterations. However, these accuracies start to decrease after iteration 9. Similarly, as observed from Figure 4.3b, there

		Computed Class			Accuracy (%)
		Normal	High-grade	Low-grade	
Actual Class	Normal	**505**	4	1	99.02
	Low-grade	3	**804**	52	93.60
	High-grade	0	12	**263**	95.64
				Overall	**95.62**

(a)

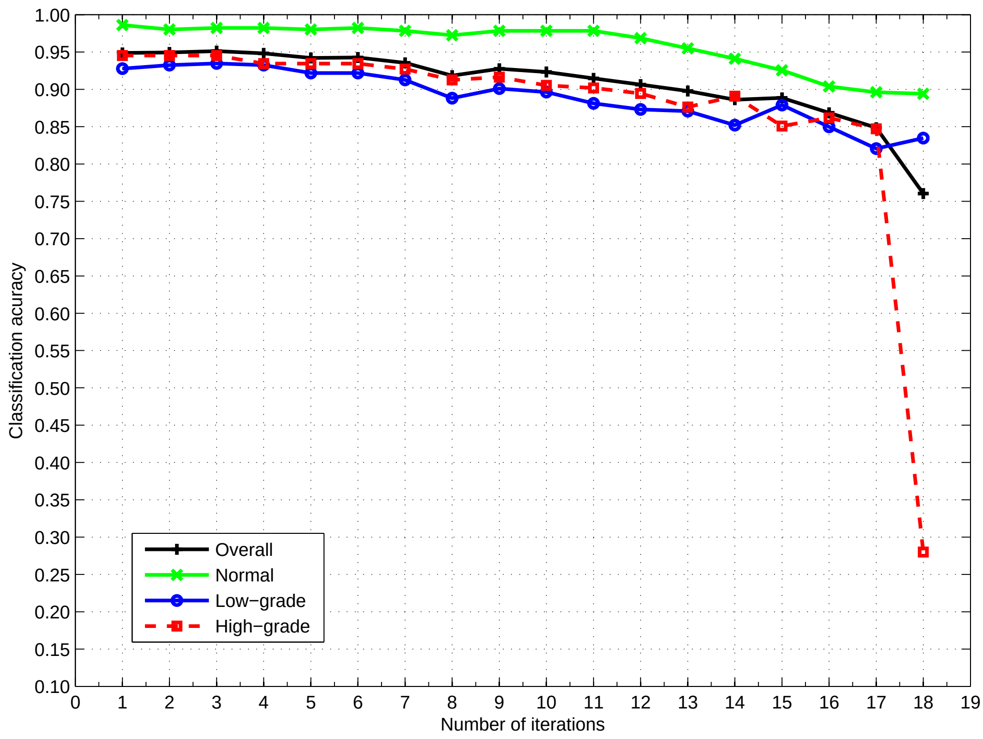
		Computed Class			Accuracy (%)
		Normal	High-grade	Low-grade	
Actual Class	Normal	**461**	23	7	93.89
	Low-grade	21	**728**	95	86.26
	High-grade	7	38	**212**	82.49
				Overall	**88.00**

(b)

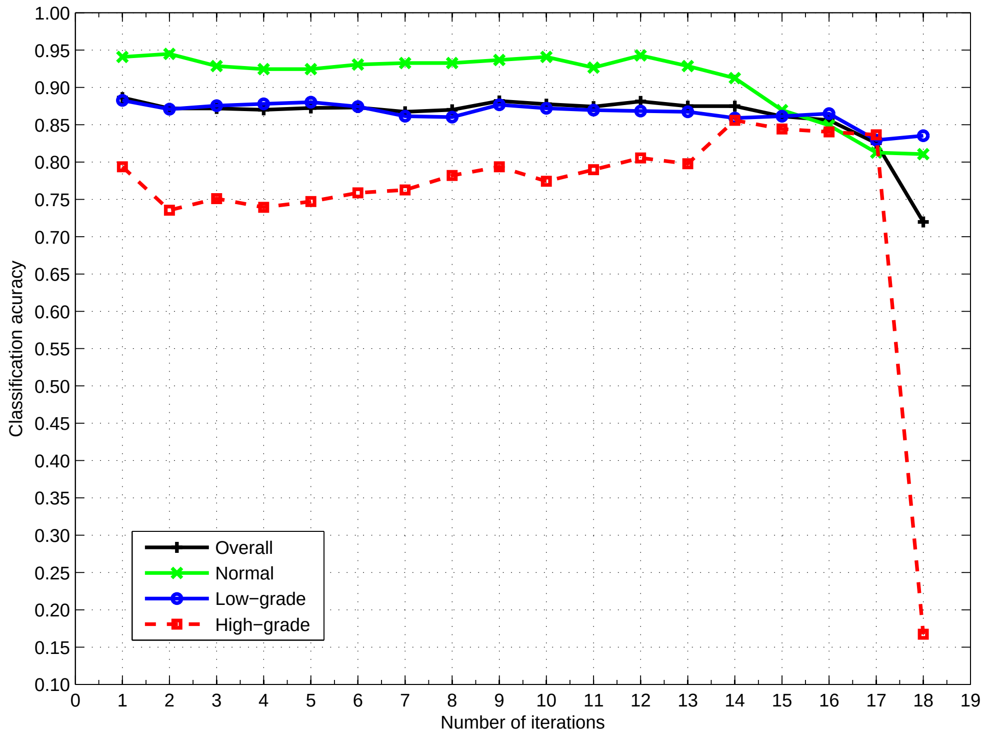
	Average	Std. dev.
Overall	88.00	0.14
Normal	93.89	0.14
Low-grade	86.26	0.18
High-grade	82.49	0.56

(c)

Table 4.3: Confusion matrices obtained with the color graph approach for (a) the training set and (b) the test set, and (c)average accuracy results and their standard deviations obtained by applying leave-one-patient-out cross validation of the test set. These results are obtained using a subset of color graph features determined by forward selection.



(a)



(b)

Figure 4.3: Classification accuracies of (a) training and (b) test sets with respect to iterations of backward elimination.

$Degree$	Average degree
$Degree(green)$	Average degree for luminal-luminal edges
$Degree(blue)$	Average degree for nuclear-nuclear edges
$Degree(magenta)$	Average degree for stromal-nuclear edges
$CC(luminal)$	Average clustering coefficient for luminal component
$Diameter$	Diameter of graph
$Diameter(green)$	Diameter for luminal-luminal edges
$Diameter(red)$	Diameter for stromal-stromal edges
$Diameter(blue)$	Diameter for nuclear-nuclear edges
$Diameter(cyan)$	Diameter for luminal-nuclear edges

Table 4.4: List of features selected by backward elimination.

is not a significant change in classification accuracies of normal and low-grade tissues in the test set until iteration 12. On the other hand, Figure 4.3b implies an increase in the classification accuracy of high-grade tissues in this interval.

Based on the results obtained on the training set, we select the iteration number as 9. The features in this subset are given in Table 4.4. Confusion matrices and leave-one-patient-out classification results that are obtained using these features are also provided in Table 4.5. Here, C parameter of support vector classifier is selected as 5 using 10-fold cross validation.

4.3 Analysis of the Method

In the previous sections, we focused our experiments on investigating the effectiveness of the proposed method. These experiments have shown that the color graph representation yields accurate classification results for both the training and test sets in the diagnosis and grading of colon cancer.

In this section, we investigate the performance of node identification that is one of the most important steps of the proposed method. Here, we focus on how robust the proposed method is to sectioning and staining related problems and how sensitive it is to the parameters that are used in the definition of circular primitives.

		Computed Class			Accuracy (%)
		Normal	High-grade	Low-grade	
Actual Class	Normal	**499**	3	8	97.84
	Low-grade	13	**774**	72	90.10
	High-grade	0	23	**252**	91.64
				Overall	**92.76**

(a)

		Computed Class			Accuracy (%)
		Normal	High-grade	Low-grade	
Actual Class	Normal	**460**	20	11	93.69
	Low-grade	27	**740**	77	87.68
	High-grade	13	40	**204**	79.38
				Overall	**88.19**

(b)

	Average	Std. dev.
Overall	88.19	0.14
Normal	93.69	0.13
Low-grade	87.68	0.17
High-grade	79.37	0.63

(c)

Table 4.5: Confusion matrices obtained with the color graph approach for (a) the training set and (b) the test set, and (c) average accuracy results and their standard deviations obtained by applying leave-one-patient-out cross validation of the test set. These results are obtained using a subset of color graph features determined by backward elimination.

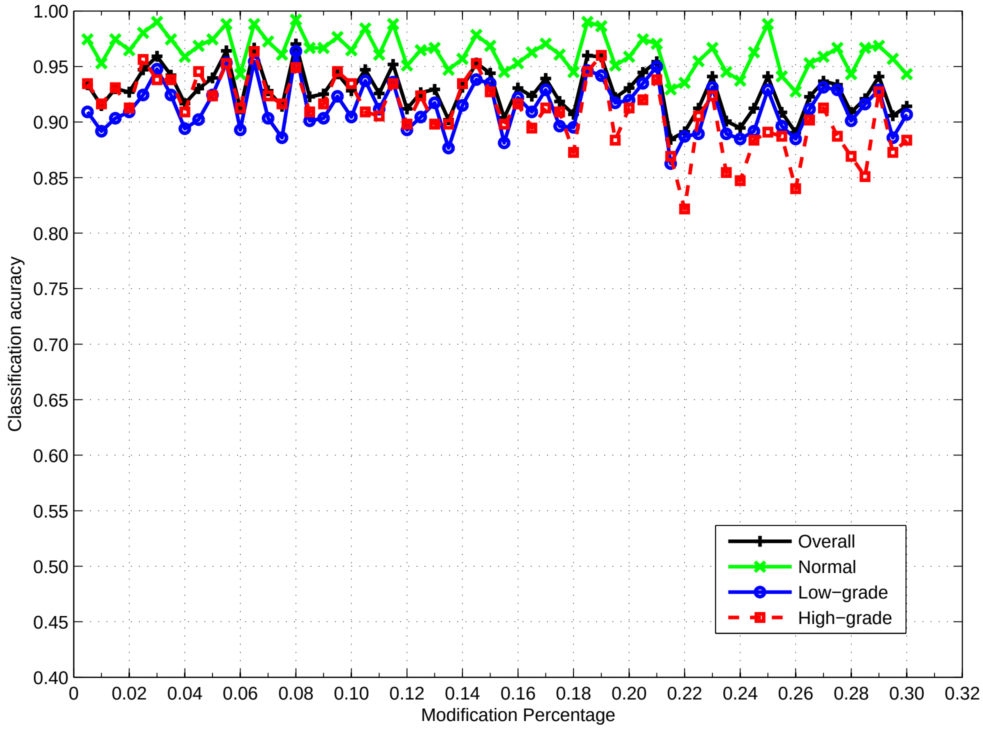
4.3.1 Analysis of Node Identification

In the node identification step, pixels of histopathological tissue images are clustered into three for unsupervised segmentation of tissue components. This segmentation mainly depends on the pixel intensity values of cytological components that are obtained with the hematoxylin-and-eosin staining technique. This technique is applied on tissue samples at different times and by different people. Thus, there could be staining differences in the staining procedure that is applied on different tissue samples. There could also exist staining problems that are caused by inappropriate application of the procedure. Therefore, pixel clusters that are obtained in this step could include mislabeled pixels which will result in incorrect segmentation of image regions belonging to different tissue components.

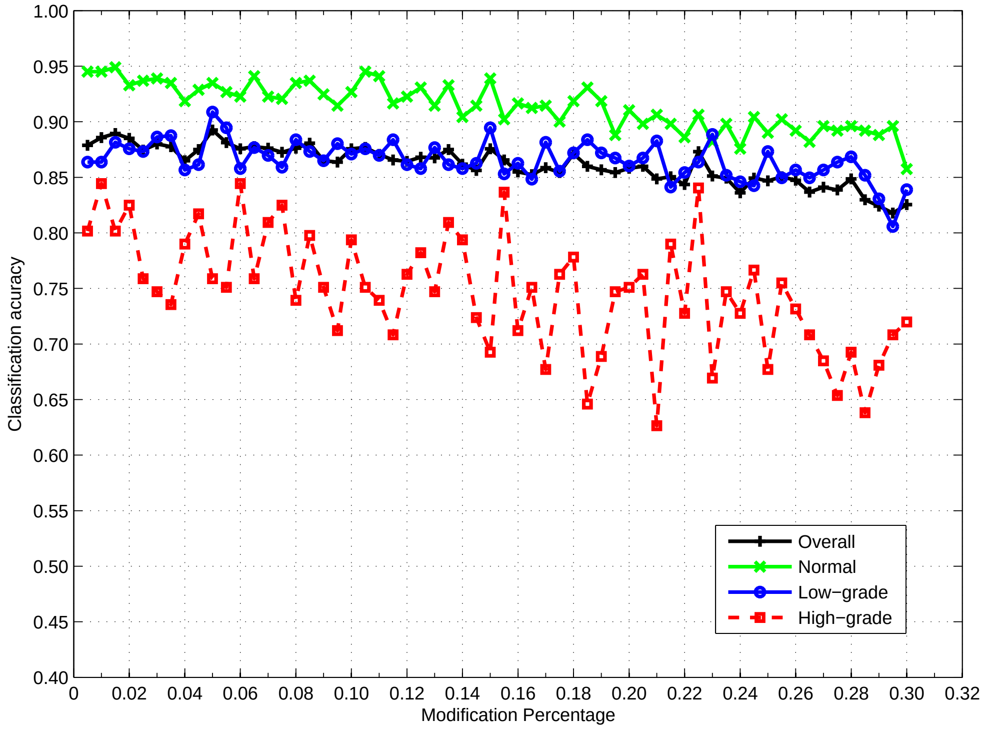
After identifying pixel regions of tissue components, the color graph method locates circular primitives on these regions. These circular primitives approximately represent particular components of a tissue, ignoring the amount of their corresponding pixels. As this representation does not locate the exact boundaries of cytological tissue components, there could be some information lost. Moreover, as there could be a considerable amount of noise arising from the tissue preparation procedures, some tissue components could be located incorrectly. Besides, the circle fitting algorithm may transform a particular tissue component to multiple circular primitives.

The aforementioned reasons may affect the effectiveness of the node identification method. Therefore, to understand the effectiveness of the proposed node identification method, we quantitatively assess the effects of incorrect identification of nodes to the classification accuracy. To this end, for a given tissue image, we make modifications on the m percent of its nodes. These modifications include a randomly selected combination of insertion, deletion, and type change operations. The definitions of these operations are as follows:

1. Insertion: This operation adds a new node into the tissue representation. The coordinates of the added node and its component type are randomly determined. For a particular tissue image, there could be sectioning and



(a)



(b)

Figure 4.4: Classification accuracies of (a) the training set and (b) the test set as a function of modification percentage.

staining related problems, such as empty regions resulted from inappropriate cutting of the tissue and existence of stain artifacts. There could also be noise in the pixel clusters that are obtained with k-means. Regarding these problems, it is possible for the proposed node identification method to generate incorrectly located nodes. The aim of the insertion operation is to simulate this kind of cases where additional nodes are generated (e.g., for the cases where a nuclear circle located on a stain artifact and luminal circles located on empty regions caused by inappropriate sectioning).

2. Deletion: This operation removes an existing node from the tissue representation. The node, which is to be removed, is randomly selected. The aim of the deletion operation is to simulate the cases where some nodes cannot be identified because of the similar reasons. An example of these cases is a nucleus that cannot be represented with a nuclear circle due to a fading.

3. Type change: This operation changes the tissue type of an existing node. The node whose type is to be changed and the new type are randomly selected. The aim of this operation is to simulate the cases where stain artifacts exist in the tissue and/or the k-means algorithm results in incorrect clustering of some pixels. An example of these cases is a cell nucleus appearing in pink because of stain artifacts.

In our experiments, we applied these three types of modifications to the nodes with varying modification percentages from the range of $[0.005, 0.3]$. Figure 4.4 shows the classification accuracies as a function of modification percentage. These accuracies are obtained using all features of color graphs.

Figure 4.4a shows that with the increasing modification percentage, the proposed method still leads to accurate results ($>$ 80 percent) for all classes in the training set. Here, the error arising from the modified nodes could most probably be compensated by the other nodes. Figure 4.4b shows that this situation is similar for normal and low-grade cancerous tissues in the test set. However, the classification accuracy of high-grade cancerous tissues in the test set decreases under 60 percent with the increasing values of the modification percentage. These results show that the proposed approach is tolerant to slight differences of locating graph nodes.

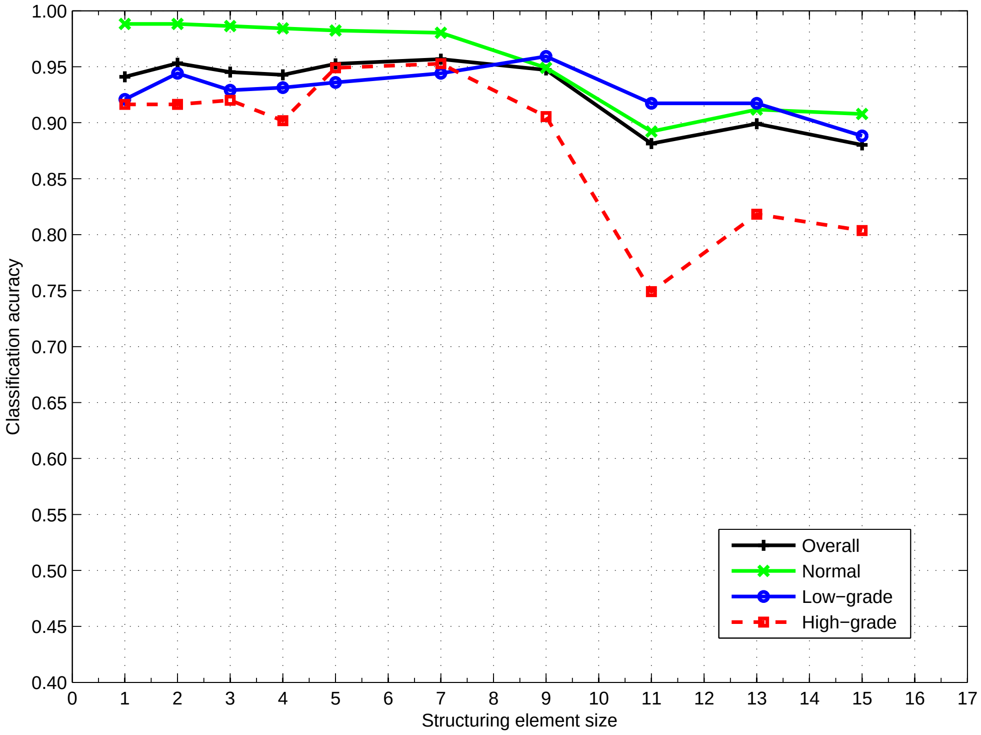
4.3.2 Analysis of Parameters

As mentioned in Section 3.1, there are two parameters of the node identification step. These are (i) *the size of the structuring element* that is used by morphological operators for reducing image noise, and (ii) *the area threshold* that is used by circle fitting for controlling the minimum area of generated circles.

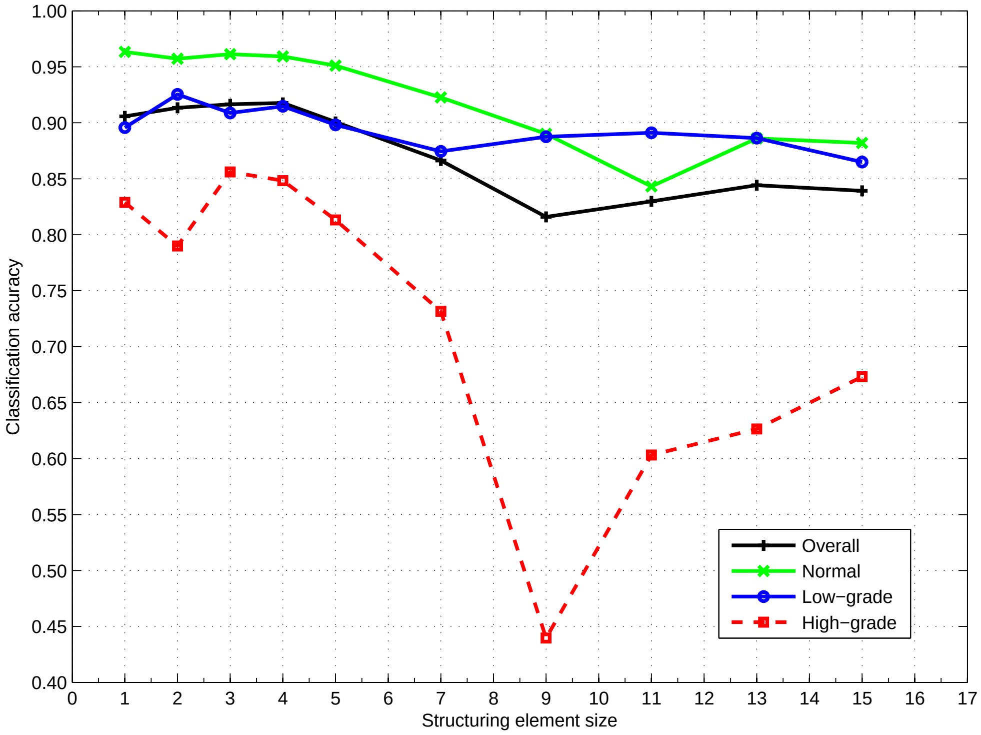
In this subsection, we aim to investigate the effects of these parameters to the classification accuracy of the color graph method. To this end, we fix one of the parameters and observe the accuracy as a function of the other parameter. In the analysis of both of these parameters, SVM parameter C is selected using 10-fold cross validation on the training set.

In Figure 4.5, we provide the classification accuracies obtained on the training set and the test set when the structuring element size is selected as $\{1, 2, 3, 4, 5, 7, 9, 11, 13, 15\}$; here the value of 1 corresponds to the case where no morphological operator is applied. Figure 4.5a shows that applying morphological operators does not improve the training accuracy, but reduces for larger structuring element sizes. On the other hand, Figure 4.5b shows that applying morphological operators improve the test accuracy. However, with the increased size of structuring element, classification accuracy of the test set decreases. As observed from the figure, applying morphological operators gives most accurate results for the sizes in between 1 and 4.

Similarly Figure 4.6 shows the classification accuracies obtained on the training set and the test set when the area threshold is selected as $\{5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 75, 100, 125, 150\}$. This figure shows that area thresholds in the range of $[5, 45]$ gives accurate results for both training and test sets. When the threshold reaches to values that are greater than 50, there is a considerable amount of decrease in the classification accuracy of high-grade

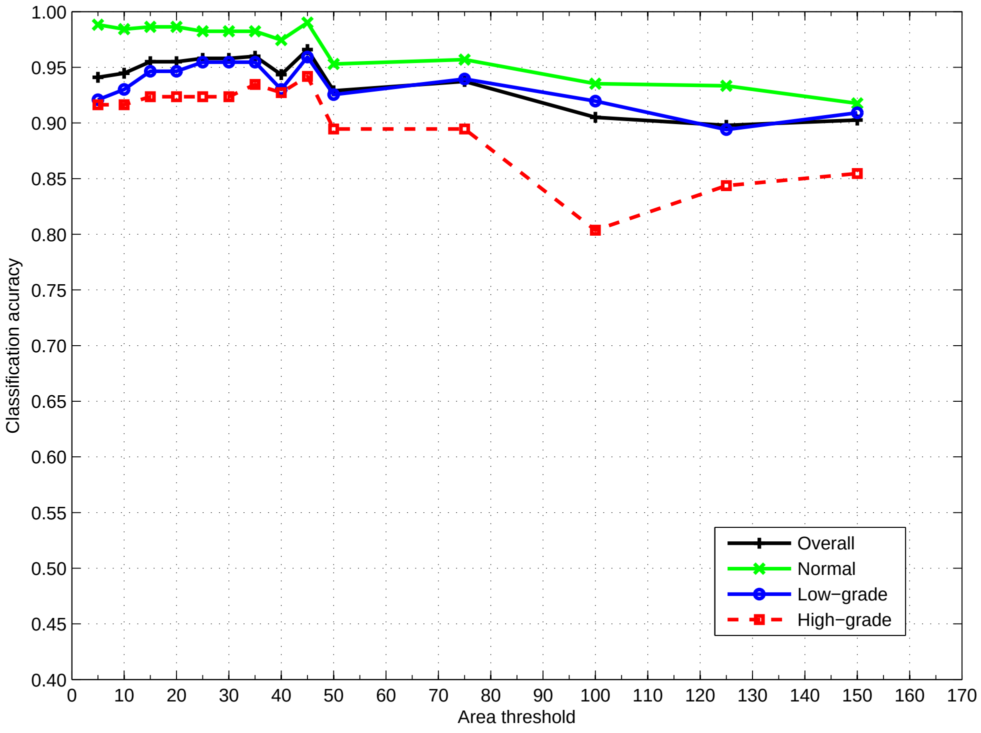


(a)

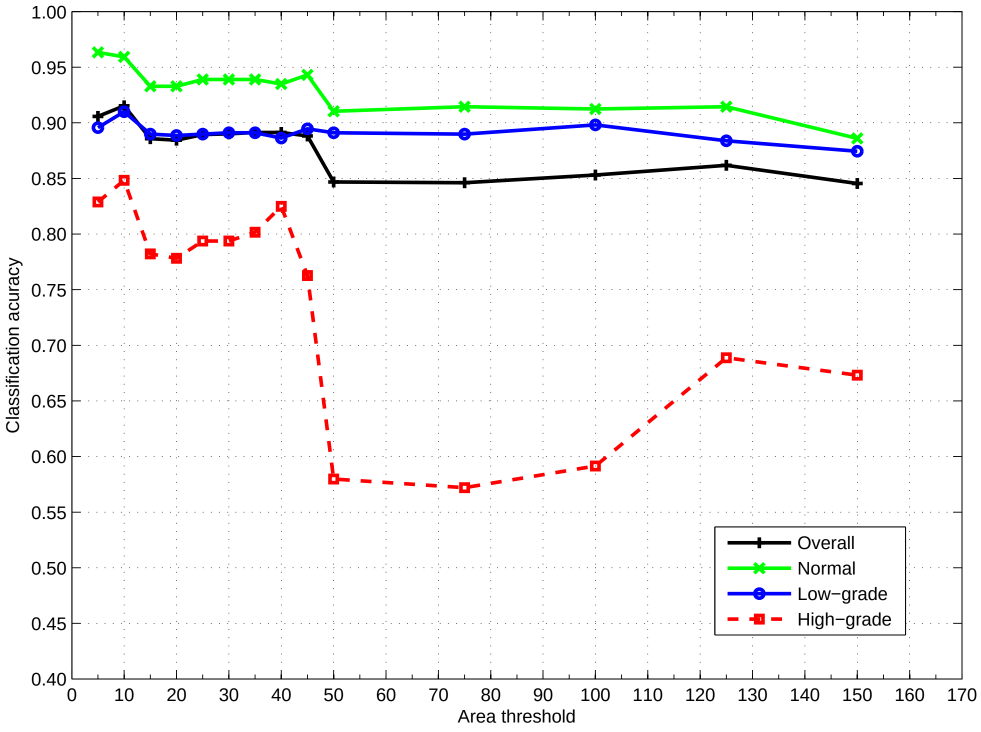


(b)

Figure 4.5: Classification accuracies of (a) the training set and (b) the test set as a function of the structuring element size.



(a)



(b)

Figure 4.6: Classification accuracies of (a) the training set and (b) the test set as a function of the area threshold, which is used in circle fitting.

cancerous tissues.

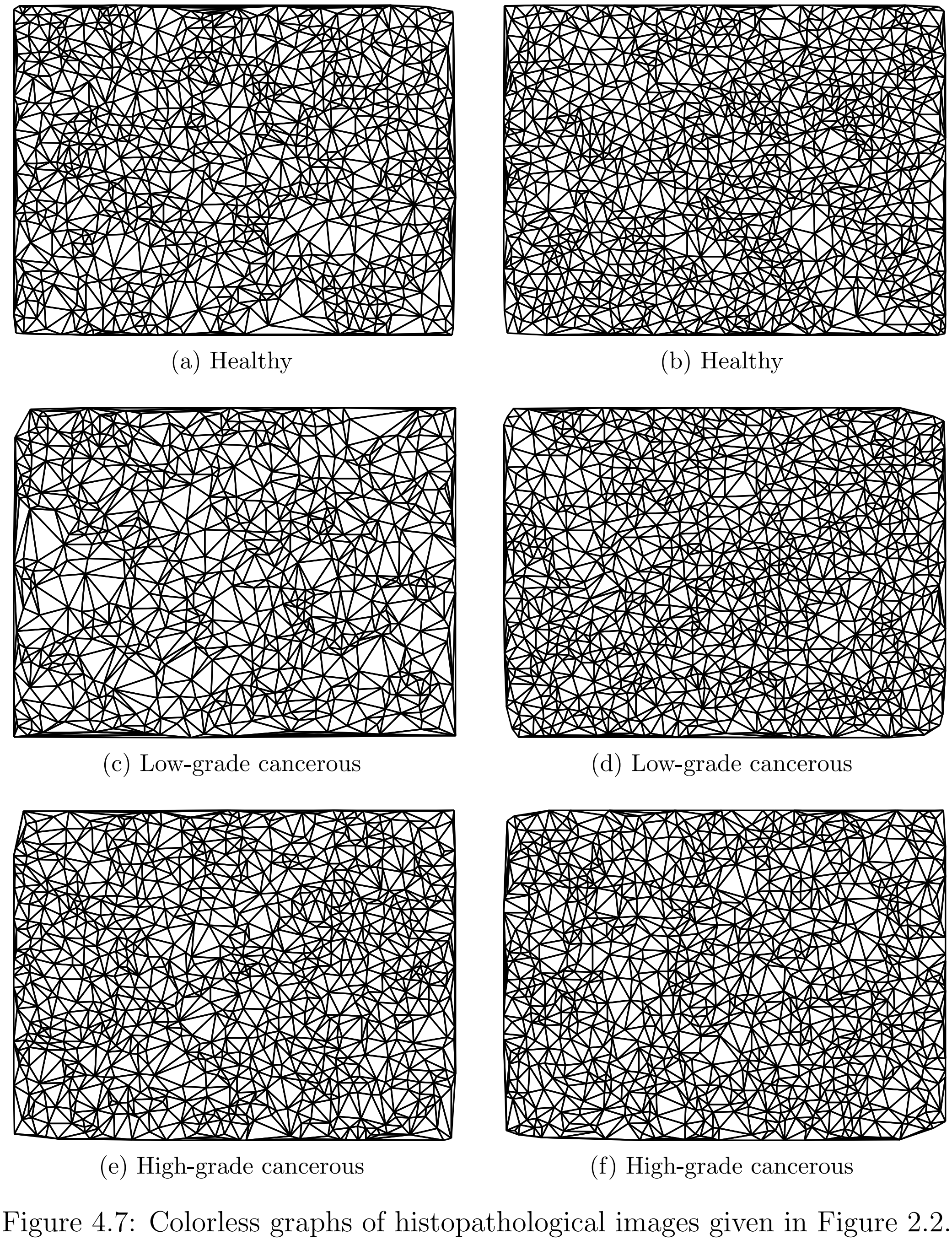
4.4 Comparisons

In this section, we provide a comparison of the proposed approach against previous structural methods to investigate the effectiveness of the color graph representation. For this reason we use features of colorless graphs and Delaunay triangulations. In all comparisons, we use support vector machines as classifiers. The details and classification results of these methods are given in the following subsections. Similarly, the SVM cost parameter C is selected from the candidate set of $\{1, 2, ..., 9, 10, 20, ...90, 100, 150, ..., 950, 1000\}$ using 10-fold cross-validation on the training set.

4.4.1 Colorless Graphs

To examine the effectiveness of features that are introduced by the color graph representation, we employ colorless graphs. Similar to the color graph method, the colorless graph representation also constructs a Delaunay triangulation considering the locations of nuclear, stromal, and luminal tissue components. However, it does not color the triangle edges. Thus, the colored versions of graph features cannot be used in this representation. We compute the average degree, average clustering coefficient, and diameter of colorless graphs to be used in classification. Figure 4.7 shows the colorless graphs that are constructed on the circles given in Figure 3.3.

In Table 4.6, for colorless graph representation, we provide the confusion matrices for the training and test sets and leave-one-patient-out cross validation results. These results show that introduction of color edges, and thus colored versions of graph features, significantly improves the classification accuracies.



(a) Healthy

(b) Healthy

(c) Low-grade cancerous

(d) Low-grade cancerous

(e) High-grade cancerous

(f) High-grade cancerous

Figure 4.7: Colorless graphs of histopathological images given in Figure 2.2.

		Computed Class			Accuracy (%)
		Normal	High-grade	Low-grade	
Actual Class	Normal	**248**	96	166	48.63
	Low-grade	252	**221**	386	25.73
	High-grade	50	29	**196**	71.27
				Overall	**40.45**

(a)

		Computed Class			Accuracy (%)
		Normal	High-grade	Low-grade	
Actual Class	Normal	**207**	104	34	42.16
	Low-grade	254	**172**	418	20.38
	High-grade	59	39	**159**	61.87
				Overall	**33.79**

(b)

	Average	Std. dev.
Overall	33.79	0.21
Normal	42.16	0.22
Low-grade	20.38	0.20
High-grade	61.86	0.58

(c)

Table 4.6: Confusion matrices obtained with the colorless graph features for (a) the training set and (b) the test set, and (c) average accuracy results and their standard deviations obtained by applying leave-one-patient-out cross validation of the test set.

4.4.2 Delaunay Triangulations

The Delaunay triangulation method only considers nuclear tissue components in structural representation. Thus, it ignores the existence of different components in the tissue. We employ this method to examine the effects of considering luminal and stromal tissue components in a structural representation. Figure 4.8 shows the Delaunay triangulations that are constructed on the nuclear circles given in Figure 3.3.

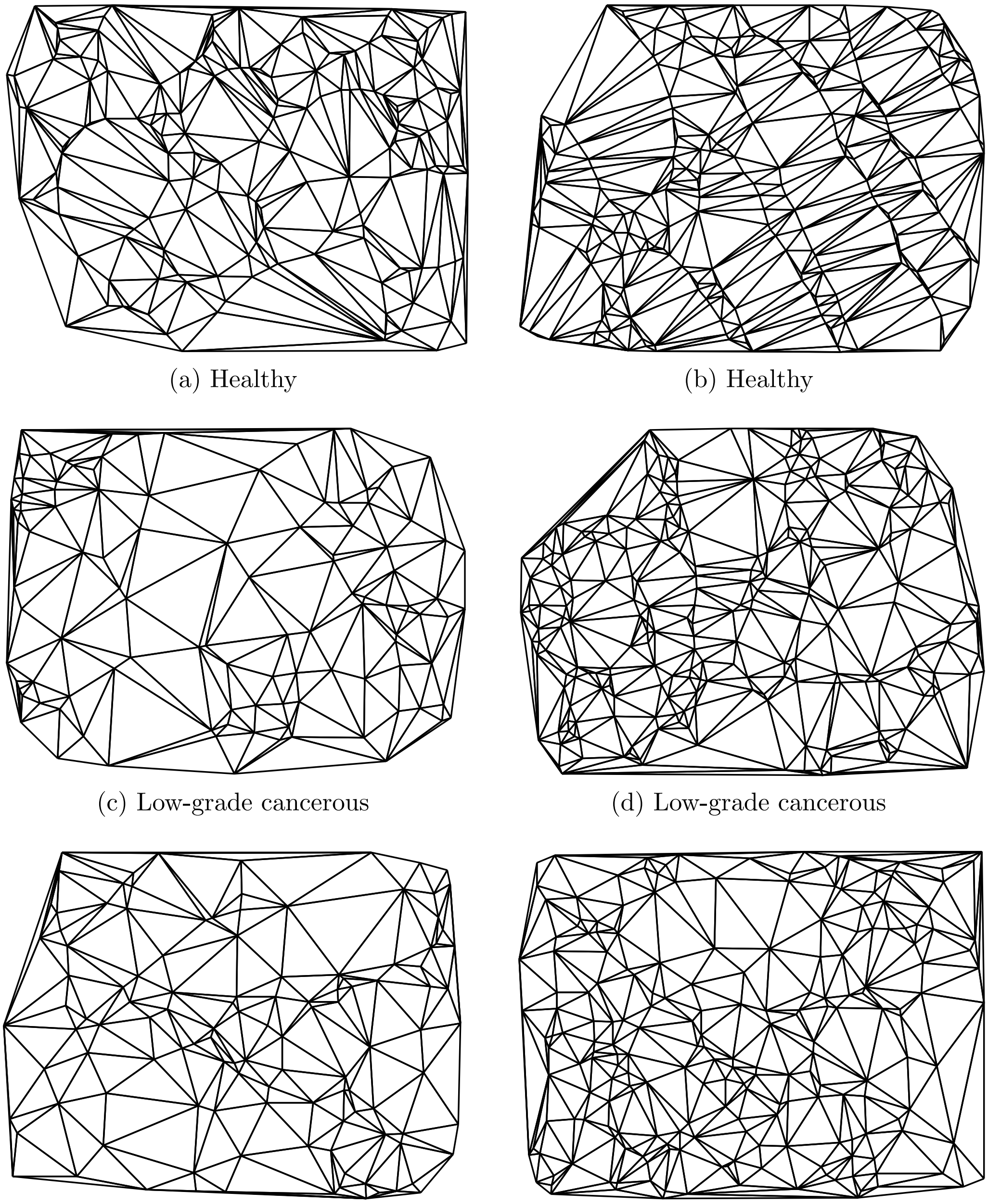
As there is no color information associated with edges, the colorless versions of the features are used to quantify this representation. To this end, the average degree, the average clustering coefficient, and the diameter are computed from the constructed graphs. Furthermore, eight more features are extracted using the areas and perimeters of the triangles [36]. For each of them, we compute the average, standard deviation, minimum-to-maximum ratio, and disorder. In the definition of disorder, we use the formula of Equation 4.1 that is given in [90].

$$disorder = 1 - \frac{1}{1 + \frac{standarddeviation}{average}} \tag{4.1}$$

As a result, a total of 11 features is used to quantify a Delaunay triangulation. Similarly, a support vector machine classifier is used and its parameter C is selected using 10-fold cross-validation. The accuracy results obtained by this Delaunay triangulation representation are given in Table 4.7. These results demonstrate that the use of the locations of additional tissue components increases the classification accuracy especially for the low-grade cancerous tissues.

4.4.3 Statistical Tests

Confusion matrices of colorless graphs and Delaunay triangulations show that the color graph approach outperforms these methods. Moreover, we employ McNemar's test to investigate how significant this improvement.



(a) Healthy

(b) Healthy

(c) Low-grade cancerous

(d) Low-grade cancerous

(e) High-grade cancerous

(f) High-grade cancerous

Figure 4.8: Delaunay triangulations of histopathological images given in Figure 2.2.

		Computed Class			Accuracy (%)
		Normal	High-grade	Low-grade	
Actual Class	Normal	**451**	46	13	88.43
	Low-grade	89	**617**	153	71.83
	High-grade	11	60	**204**	74.18
				Overall	**77.37**

(a)

		Computed Class			Accuracy (%)
		Normal	High-grade	Low-grade	
Actual Class	Normal	**433**	34	24	88.19
	Low-grade	94	**610**	140	72.27
	High-grade	27	48	**182**	70.82
				Overall	**76.95**

(b)

	Average	Std. dev.
Overall	76.95	0.15
Normal	88.19	0.19
Low-grade	72.28	0.22
High-grade	70.81	0.45

(c)

Table 4.7: Confusion matrices obtained with the Delaunay triangulation features for (a) the training set and (b) the test set, and (c) average accuracy results and their standard deviations obtained by applying leave-one-patient-out cross validation of the test set.

The McNemar's test is a non-parametric statistical method proposed by Quinn McNemar for testing the marginal homogeneity of 2 × 2 contingency tables [72]. This method is used to compare two population proportions that are correlated or dependent to each other. McNemar's test is also used for testing the difference between two raters, that is the case in this study.

		Rater 2	
		+	-
Rater 1	+	A	B
	-	C	D

Table 4.8: 2 × 2 contingency matrix of McNemar's test

Table 4.8 shows a 2 × 2 contingency table representing the pairwise results of two raters for a positive/negative classification problem. The null hypothesis of the McNemar's test claims that row totals of this table are equal to column totals, such that

$$(A + B) = (A + C) \;\text{ and } (B + D) = (C + D) \tag{4.2}$$

If we eliminate A and D in Equation 4.2, we obtain $B = C$ as the null hypothesis which means that there is no significant difference between the decisions of these raters. McNemar's test uses a χ^2 statistics for testing this hypothesis which is computed using Equation 4.3.

$$\chi^2 = \frac{(B - C)^2}{B + C} \tag{4.3}$$

For testing the null hypothesis in McNemar's test, this value is compared with the χ^2 table values with 1 degree of freedom. If the calculated χ^2 value is greater than the table value, the null hypothesis is rejected. Otherwise, the null hypothesis is accepted, which means that there is not a significant difference between the decisions of the raters. Moreover, there exist a P value for each element of χ^2 table which indicates the significance level of this decision.

McNemar's test is designed for testing marginal homogeneity of 2×2 contingency tables. However, one could make use of McNemar's test on $N \times N$ tables by collapsing them to 2×2 tables using a one-against-all manner. This way, it is possible to obtain N individual tests of 2×2 tables corresponding to each level of the table.

In this study, we employ McNemar's test for investigating how significant the differences between color graphs and its counterparts are. To this end, we compute χ^2 statistics and their corresponding P values. Here, we use the contingency tables given in Table 4.9.

		Colorless Graph			
		Normal	High-grade	Low-grade	Total
Color Graph	Normal	209	103	188	500
	Low-grade	248	166	386	800
	High-grade	63	46	183	292
	Total	520	315	757	1592

(a)

		Delaunay Triangulation			
		Normal	High-grade	Low-grade	Total
Color Graph	Normal	429	41	30	500
	Low-grade	92	579	129	800
	High-grade	33	72	187	292
	Total	554	692	346	1592

(b)

Table 4.9: 3×3 contingency tables for (a) color graphs vs. colorless graphs, and (b) color graphs vs. Delaunay triangulations.

	Colorless Graph		Delaunay Triangulation	
	χ^2	P	χ^2	P
Normal	0.664	0.4250	14.878	0.0001
Low-grade	300.415	0.0000[1]	34.922	0.0000[1]
High-grade	316.581	0.0000[1]	11.045	0.0009

Table 4.10: Computed χ^2 values and their corresponding P values for colorless graphs and Delaunay triangulations.

[1] These P values are smaller than 0.0001, and thus, the exact values are not shown in the table.

Table 4.10 shows computed χ^2 values and their corresponding P values that are obtained by employing McNemar's test on the tables given in Table 4.9. These values show that with a significance level of 0.001, the classification results of the color graph approach and those of the Delaunay triangulation method is different. Hence, the improvement achieved by the color graph approach is significant. Table 4.10 also shows that, the improvement of the color graph approach on the colorless graph method is significant with a level of 0.001 for low-grade and high-grade samples. However, there is not a difference between the decisions of these methods for normal tissues with a significance level of 0.05.

Chapter 5

Conclusion and Discussion

Computer aided image analysis tools for automated cancer diagnosis and grading are gaining importance in medicine as they provide objective mathematical measures. These tools make use of computational methods that employ morphological, textural, and structural image features. Numerous studies in this field use such features of cell nuclei in order to quantify histopathological images of tissues. However, these methods ignore the existence of other components in a tissue such as luminal and stromal regions. Besides, these components play important roles in forming the tissue structure. This issue gains more importance in the cases where tissues have hierarchical structures.

In this thesis, we introduce a novel structural method for automated cancer diagnosis and grading. This method proposes to model spatial relationships of different tissue components. For this purpose, it first clusters the pixels of histopathological tissue images into three and locates circles on these pixel groups to approximately represent different cytological tissue components. Then, it constructs a Delaunay triangulation on these circular primitives, that are identified as graph nodes, and colors the triangle edges according to the component types of their end nodes. On the generated color graphs, a new set of structural features is defined considering the color information. These new features are used by support vector machines for classification.

We conduct the experiments on 3236 photomicrographs of colon tissues that are taken from 258 different patients. These experiments demonstrate that the colored versions of features lead to 94.89 percent training accuracy and 88.63 percent test accuracy for the automated colon cancer diagnosis and grading. In this thesis, we also compare the color graph approach with the colorless graph and Delaunay triangulation methods. Our experiments show that considering different tissue components in a structural representation improves the effectiveness of the method, and that introduction of color information into the graph features contributes to this improvement. McNemar's tests applied on these results show that this improvement is statistically significant.

For generating a graph based representation, we propose to use Delaunay triangulation since it is known to be one of the effective representations in quantifying the spatial distribution of graph nodes and since it does not require selecting an external parameter, for example, as in the case of probabilistic graphs in which parameter α should be selected. However, it is possible to employ alternative graph generation methods such as probabilistic graphs and Gabriel's graphs. Similarly, edges of these alternative graphs could be colored for employing the proposed color features for the quantification of tissues. Furthermore, Voronoi diagrams could be used for constructing a graph based representation. Here, a Voronoi polygon and/or its edges could be colored according to the type of its component. Similarly, the edges of polygons could be colored according to the types of their neighboring components. It is also possible to define colored versions of Voronoi diagrams. All these aspects could be considered as the future research perspectives of using the coloring idea.

Nodes of the constructed color graphs are circular primitives that approximately represent individual tissue components. As previously mentioned, the circle fitting method could split a tissue component into multiple circles resulting in a number of nodes that is relatively larger than the actual number of tissue components. To avoid this difference, constructed color graphs could be simplified by merging the nodes of the same type according to their neighborhood properties. For example, a particular luminal node having only luminal node neighbors could be merged with its neighbors to construct a "super node" that represents a group of luminal nodes. Similarly, color clustering coefficients could be used for simplifying the generated graphs. On the other hand, one could use elliptical objects rather than using circles to achieve a more effective localization of individual tissue components. However, locating ellipses is expected to have longer computational times.

Bibliography

[1] Cancer facts and figures 2009. Technical report, American Cancer Society, Atlanta, GA, 2009.

[2] Cancer prevention and early detection facts and figures 2009. Technical report, American Cancer Society, Atlanta, GA, 2009.

[3] A. N. Akansu and R. A. Haddad. *Multiresolution Signal Decomposition: Transforms, Subbands, and Wavelets*. Academic Press, Inc., Orlando, FL, USA, 1992.

[4] F. Albregtsen, B. Nielsen, and H. E. Danielsen. Adaptive gray level run length features from class distance matrices. *Pattern Recognition, International Conference on*, 3:3746, 2000.

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

[6] N. H. Anderson, P. W. Hamilton, P. H. Bartels, D. Thompson, R. Montironi, and J. M. Sloan. Computerized scene segmentation for the discrimination of architectural features in ductal proliferative lesions of the breast. *The Journal of Pathology*, 181(4):374–380, 1997.

[7] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies. Image coding using wavelet transform. *IEEE Transactions on Image Processing*, 1(2):205–220, 1992.

[8] J. W. Baish and R. K. Jain. Fractals and cancer. *Cancer Research*, 60(14):3683–3688, 2000.

[9] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, 1996.

[10] M. Bibbo, F. Michelassi, P. H. Bartels, H. Dytch, C. Bania, E. Lerma, and A. G. Montag. Karyometric marker features in normal-appearing glands adjacent to human colonic adenocarcinoma. *Cancer Research*, 50(1):147–151, 1990.

[11] C. Bilgin, C. Demir, C. Nagi, and B. Yener. Cell-graph mining for breast tissue modeling and classification. *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 5311–5314, 2007.

[12] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992.

[13] A. Cetera. The minkowski dimension and critical effects in fractal evolution of defects. *Chaos, Solitons & Fractals*, 12(3):475–482, 2001.

[14] W. Chan and K. H. Fu. Value of routine histopathological examination of appendices in Hong Kong. *Journal of Clinical Pathology*, 40(4):429–433, 1987.

[15] S. G. Chang, B. Yu, and M. Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing*, 9(9):1532–1546, 2000.

[16] C. Chen, J. DaPonte, and M. Fox. Fractal feature analysis and classification in medical imaging. *Medical Imaging, IEEE Transactions on*, 8(2):133 –142, jun 1989.

[17] H. K. Choi, T. Jarkrans, E. Bengtsson, J. Vasko, K. Wester, P-U Malmstrom, and C. Busch. Image analysis based grading of bladder carcinoma.

comparison of object, texture and graph based methods and their reproducibility. *Analytical Cellular Pathology*, 15(1):1–18, 1997.

[18] A. Chu, C. Sehgal, and J. Greenleaf. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters*, 11:415–420, 1990.

[19] D. Comaniciu, P. Meer, and D. J. Foran. Image-guided decision support system for pathology. *Mach. Vision Appl.*, 11(4):213–224, 1999.

[20] I. S. Cook and C. E. Fuller. Does histopathological examination of breast reduction specimens affect patient management and clinical follow up? *Journal of Clinical Pathology*, 57(3):286–289, 2004.

[21] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[22] K. Crammer, Y. Singer, N. Cristianini, J. Shawe-taylor, and B. Williamson. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:2001, 2001.

[23] E. A. B. da Silva, D. G. Sampson, and M. Ghanbari. A successive approximation vector quantizer for wavelet transform image coding. *IEEE Transactions on Image Processing*, 5(2):299–310, 1996.

[24] B. V. Dasarathy and E. B. Holder. Image characterizations based on joint gray level-run length distributions. *Pattern Recogn. Lett.*, 12(8):497–502, 1991.

[25] I. Daubechies, editor. *Different Perspectives on Wavelets.* American Mathematical Society, 1992.

[26] I. Daubechies. *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[27] G. V. de Wouwer, B. Weyn, P. Scheunders, W. Jacob, E. V. Marck, and D. V. Dyck. Wavelets as chromatin texture descriptors for the automated identification of neoplastic nuclei. *Journal of Microscopy*, 197(1):25–35, 2000.

[28] B. Delaunay. Sur la sphere vide. a la memoire de georges voronoi. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskih i Estestvennyh Nauk*, 7:793–800, 1934.

[29] C. Demir, S. H. Gultekin, and B. Yener. Augmented cell-graphs for automated cancer diagnosis. *Bioinformatics*, 21(2):7–12, 2005.

[30] C. Demir, S. H. Gultekin, and B. Yener. Learning the topological properties of brain tumors. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(3):262–270, 2005.

[31] R. Descartes. *Principia Philosophiae.* Ludovicus Elzevirius, Amsterdam, 1644.

[32] G. L. Dirichlet. über die reduktion der positiven quadratischen formen mit drei unbestimmten ganzen zahlen. *Journal für die Reine und Angewandte Mathematik*, 40:209–227, 1850.

[33] M. F. Dixon, L. J. R. Brown, H. M. Gilmour, A. B. Price, N. C. Smeeton, I. C. Talbot, and G. T. Williams. Observer variation in the assessment of dysplasia in ulcerative colitis. *Histopathology*, 13(4):385–397, 1988.

[34] R. Dobrescu, F. Talos, and C. Vasilescu. Using fractal dimension for cancer diagnosis. *Video/Image Processing and Multimedia Communications 4th EURASIP-IEEE Region 8 International Symposium on VIPromCom*, pages 173–176, 2002.

[35] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. In *Adv. Phys*, pages 1079–1187, 2002.

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

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

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

[39] A. N. Esgiar, R. N. G. Naguib, B. S. Sharif, M. K. Bennett, and A. Murray. Fractal analysis in the detection of colonic cancer images. *IEEE Transactions on Information Technology in Biomedicine*, 6(1):54–58, 2002.

[40] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller. Hematoxylin and eosin staining of tissue and cell sections. *Cold Spring Harbor Protocols*, 2008(6), 2008.

[41] P. Fleischmann. *Mesh generation for technology CAD in three dimensions.* Dissertation, Institute for Microelectronics, Technical University Vienna, Austria, 1999.

[42] S. Fortune. A sweepline algorithm for voronoi diagrams. In *SCG '86: Proceedings of the second annual symposium on Computational geometry*, pages 313–322, New York, NY, USA, 1986. ACM.

[43] S. Fortune. Stable maintenance of point set triangulations in two dimensions. *Foundations of Computer Science, Annual IEEE Symposium on*, 0:494–499, 1989.

[44] K. R. Gabriel and R. R. Sokal. A new statistical approach to geographic variation analysis. *Systematic Zoology*, 18:259–278, 1969.

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

[46] M. Garcia, A. Jemal, E. M. Ward, M. M. Center, Y. Hao, R. L. Siegel, and M. J. Thun. Global cancer facts and figures 2007. Technical report, American Cancer Society, Atlanta, GA, 2007.

[47] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, 2001.

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

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

[50] S. Grgic, M. Grgic, and B. Zovko-Cihlar. Performance analysis of image compression using wavelets. *IEEE Transactions on Industrial Electronics*, 48(3):682–695, 2001.

[51] L. J. Guibas, D. E. Knuth, and M. Sharir. Randomized incremental construction of delaunay and voronoi diagrams. *Algorithmica*, 7(1):381–413, June 1992.

[52] M. Guillaud, D. Cox, K. Adler-Storthz, A. Malpica, G. Staerkel, J. Matisic, D. Van Niekerk, N. Poulin, M. Follen, and C. MacAulay. Exploratory analysis of quantitative histopathology of cervical intraepithelial neoplasia: Objectivity, reproducibility, malignancy-associated changes, and human papillomavirus. *Cytometry*, 60A(1):81–9, 2004.

[53] C. Gunduz, B. Yener, and S. H. Gultekin. The cell graphs of cancer. *Bioinformatics*, 20(1):145–151, 2004.

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

[55] C. Gunduz-Demir, M. Kandemir, A. B. Tosun, and C. Sokmensuer. Automatic segmentation of colon glands using object-graphs. *Medical Image Analysis*, 14(1):1 – 12, 2010.

[56] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.

[57] P. W. Hamilton, P. H. Bartels, D. Thompson, N. H. Anderson, R. Montironi, and J. M. Sloan. Automated location of dysplastic fields in colorectal histology using image texture analysis. *The Journal of Pathology*, 182(1):68–75, 1997.

[58] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621, 1973.

[59] D. R. Hinton, E. Dolan, and A. A. Sima. The value of histopathological examination of surgically removed blood clot in determining the etiology of spontaneous intracerebral hemorrhage. *Stroke*, 15(3):517–520, 1984.

[60] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415 –425, mar 2002.

[61] R. S. H. Istepanian and A. A. Petrosian. Optimal zonal wavelet-based ECG data compression for a mobile telecardiology system. *IEEE Transactions on Information Technology in Biomedicine*, 4(3):200–211, 2000.

[62] K. Jafari-Khouzani and H. Soltanian-Zadeh. Multiwavelet grading of pathological images of prostate. *IEEE Transactions on Biomedical Engineering*, 50(6):697–704, 2003.

[63] A. Jemal, R. Siegel, and E. Ward. Colorectal cancer facts and figures 2008-2010. Technical report, American Cancer Society, Atlanta, GA, 2008.

[64] A. E. Jones, A. W. Phillips, J. R. Jarvis, and K. Sargen. The value of routine histopathological examination of appendicectomy specimens. *BMC Surgery*, 7:17, 2007.

[65] S. J. Keenan, J. Diamond, G. W. McCluggage, H. Bharucha, D. Thompson, P. H. Bartels, and P. W. Hamilton. An automated machine vision system

for the histological grading of cervical intraepithelial neoplasia (CIN). *The Journal of Pathology*, 192(3):351–362, 2000.

[66] U. H. Kressel. Pairwise classification and support vector machines. pages 255–268, 1999.

[67] P. Kube and A. Pentland. On the imaging of fractal surfaces. *Patern Analysis and Machine Intelligence, IEEE Transactions on*, 10(5):704–707, sep 1988.

[68] R. D. Lillie. *Histopathologic Technic and Practical Histochemistry.* McGraw-Hill, 1965.

[69] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. on Knowl. and Data Eng.*, 17(4):491–502, 2005.

[70] J. B. Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[71] B. B. Mandelbrot. *The Fractal Geometry of Nature.* W H Freedman and Co., New York, 1983.

[72] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, June 1947.

[73] M. K. Mihcak, I. Kozintsev, K. Ramchandran, and P. Moulin. Low-complexity image denoising based on statistical modeling of wavelet coefficients. *IEEE Signal Processing Letters*, 6(12):300–303, 1999.

[74] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu. *Spatial tessellations: Concepts and applications of Voronoi diagrams.* Probability and Statistics. Wiley, NYC, 2nd edition, 2000.

[75] A. P. Pentland. Fractal-based description of natural scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(6):661 –674, nov. 1984.

[76] R. M. Rangayyan, L. Shen, Y. Shen, J. E. L. Desautels, H. Bryant, T. J. Terry, N. Horeczko, and M. S. Rose. Improvement of sensitivity of breast cancer diagnosis with adaptive neighborhood contrast enhancement of mammograms. *IEEE Transactions on Information Technology in Biomedicine*, 1(3):161–170, 1997.

[77] S. V. Rao. *Some studies on Beta-Skeletons.* Dissertation, Kanpur Indian Institute of Technology, 1998.

[78] J. R. Sack and J. Urrutia. *Handbook of computational geometry.* North-Holland Publishing Co., Amsterdam, The Netherlands, 2000.

[79] S. Santoso, E. J. Powers, and W. M. Grady. Power quality disturbance data compression using wavelet transform methods. *IEEE Transactions on Power Delivery*, 12(3):1250–1257, 1997.

[80] G. Schaller and M. Meyer-Hermann. Multicellular tumor spheroid in an off-lattice Voronoi-Delaunay cell model. *Phys. Rev. E*, 71(5):051910, May 2005.

[81] A. Schatzkin, L. S. Freedman, S. M. Dawsey, and E. Lanza. Interpreting precursor studies: what polyp trials tell us about large-bowel cancer. *Journal of the National Cancer Institute*, 86(14)(1053):7, July 1994.

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

[83] J. K. Shuttleworth, A. G. Todman, R. N. G. Naguib, B. M. Newman, and M. K. Bennett. Colour texture analysis using co-occurrence matrices for classification of colon cancer images. *Canadian Conference on Electrical and Computer Engineering*, 2:1134–1139, 2002.

[84] J. K. Shuttleworth, A. G. Todman, R. N. G. Naguib, B. M. Newman, and M. K. Bennett. Multiresolution colour texture analysis for classifying

colon cancer images. *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, 2:1118–1119, 2002.

[85] A. C. Society, U. M.-S. T. F. on Colorectal Cancer, and A. C. of Radiology. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2007: A joint guideline from the american cancer society, the u.s. multi-society task force on colorectal cancer, and the american college of radiology. *CA Cancer Journal for Clinicians*, 58(3):130–160, 2008.

[86] P. Spyridonos, P. Ravazoula, D. Cavouras, K. Berberidis, and G. Nikiforidis. Computer-based grading of haematoxylin-eosin stained tissue sections of urinary bladder carcinomas. In *Inform. Internet Med*, pages 179–190, 2001.

[87] S. L. Stewart, J. M. Wike, I. Kato, D. R. Lewis, and F. Michaud. A population-based study of colorectal cancer histology in the united states. *CA Cancer Journal for Clinicians*, 107(S5):1128–1141, 2006.

[88] W. N. Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Proc. Int. Symp. on Electronic Imaging: Science and Technology*, volume 1905, pages 861–870, 1993.

[89] V. Strela, P. N. Heller, G. Strang, P. Topiwala, and C. Heil. The application of multiwavelet filterbanks to image processing. *IEEE Transactions on Image Processing*, 8(4):548–563, 1999.

[90] 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*, 21:71–86, 2000.

[91] G. Sun, X. Dong, and G. Xu. Tumor tissue identification based on gene expression data using DWT feature extraction and pnn classifier. *Neurocomputing*, 69(4-6):387–402, 2006.

[92] X. Tang. Texture information in run-length matrices. *IEEE Transactions on Image Processing*, 7(11):1602–1609, 1998.

[93] G. D. Thomas, M. F. Dixon, N. C. Smeeton, and N. S. Williams. Observer variation in the histological grading of rectal carcinoma. *Journal of Clinical Pathology*, 36(4):385–391, 1983.

[94] A. B. 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.

[95] F. Truchetet and O. Laligant. Industrial applications of the wavelet and multi-resolution-based signal/image processing: a review. In D. Fofi and F. Meriaudeau, editors, *Eighth International Conference on Quality Control by Artificial Vision*, volume 6356. SPIE, 2007.

[96] V. N. Vapnik. *Statistical Learning Theory.* Wiley-Interscience, 1998.

[97] G. Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die Reine und Angewandte Mathematik*, 133:97–178, 1907.

[98] G. Voronoi. Deuxiéme mémoire: recherches sur les paralléloedres primitifs. *Journal für die Reine und Angewandte Mathematik*, 136:67–181, 1909.

[99] R. Walker, P. Jackway, B. Lovell, and I. Longstaff. Classification of cervical cell nuclei using morphological segmentation and textural feature extraction. In *Intelligent Information Systems,1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pages 297 –301, 29 1994.

[100] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *In Proceedings of the Seventh European Symposium On Artificial Neural Networks*, 1999.

[101] B. Weyn, G. V. de Wouwer, A. V. Daele, P. Scheunders, D. V. Dyck, E. V. Marck, and W. Jacob. Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. *Cytometry*, 33(1):32–40, 1998.

[102] B. Weyn, G. V. de Wouwer, G. K. Samir, A. V. Daele, P. Scheunders, E. V. Marck, and W. Jacob. Computer-assisted differential diagnosis of

malignant mesothelioma based on syntactic structure analysis. *Cytometry*, 35:23–29, 1999.

[103] B. Weyn, W. Jacob, V. D. da Silva, R. Montironi, P. W. Hamilton, D. Thompson, H. G. Bartels, A. V. Daele, K. Dillon, and P. H. Bartels. Data representation and reduction for chromatin texture in nuclei from premalignant prostatic, esophageal, and colonic lesions. *Cytometry*, 41(3):133–138, 2000.

[104] B. Weyn, G. van de Wouver, M. Koprowski, A. van Daele, K. Dhaene, P. Scheunders, W. Jacob, and E. van Marck. Value of morphometry, texture analysis, densitometry, and histometry in the differential diagnosis and prognosis of malignant mesothelia. *The Journal of Pathology*, 189(4):581–589, 1999.

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

[106] W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology*, 26, 1995.