CONSTRAINED DELAUNAY TRIANGULATION FOR DIAGNOSIS AND GRADING OF COLON CANCER

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

Süleyman Tuncer Erdoğan

July, 2009

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. Pınar Şenkul

Approved for the Institute of Engineering and Science:

Prof. Dr. Mehmet B. Baray
Director of the Institute

ABSTRACT

CONSTRAINED DELAUNAY TRIANGULATION FOR DIAGNOSIS AND GRADING OF COLON CANCER

Süleyman Tuncer Erdoğan
M.S. in Computer Engineering
Supervisor: Assist. Prof. Dr. Çiğdem Gündüz Demir
July, 2009

In our century, the increasing rate of cancer incidents makes it inevitable to employ computerized tools that aim to help pathologists more accurately diagnose and grade cancerous tissues. These mathematical tools offer more stable and objective frameworks, which cause a reduced rate of intra- and inter-observer variability. There has been a large set of studies on the subject of automated cancer diagnosis/grading, especially based on textural and/or structural tissue analysis. Although the previous structural approaches show promising results for different types of tissues, they are still unable to make use of the potential information that is provided by tissue components rather than cell nuclei. However, this additional information is one of the major information sources for the tissue types with differentiated components including luminal regions being useful to describe glands in a colon tissue.

This thesis introduces a novel structural approach, a new type of constrained Delaunay triangulation, for the utilization of non-nuclei tissue components. This structural approach first defines two sets of nodes on cell nuclei and luminal regions. It then constructs a constrained Delaunay triangulation on the nucleus nodes with the lumen nodes forming its constraints. Finally, it classifies the tissue samples using the features extracted from this newly introduced constrained Delaunay triangulation.

Working with 213 colon tissues taken from 58 patients, our experiments demonstrate that the constrained Delaunay triangulation approach leads to higher accuracies of 87.83 percent and 85.71 percent for the training and test sets, respectively. The experiments also show that the introduction of this new structural representation, which allows definition of new features, provides a more robust graph-based methodology for the examination of cancerous tissues and

better performance than its predecessors.

Keywords: Constrained Delaunay triangulation, histopathological image analysis, automated cancer diagnosis and grading, colon cancer, adenocarcinoma.

ÖZET

KOLON KANSERİNİN KISITLI DELAUNAY ÜÇGENLEMESİ İLE TEŞHİSİ VE SINIFLANDIRILMASI

Süleyman Tuncer Erdoğan
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Y. Doç. Dr. Çiğdem Gündüz Demir
Temmuz, 2009

Yüzyılımızda artan kanser vakaları, bilgisayar destekli araçların kullanımını kaçınılmaz kılmıştır; bunlar patologların kanserli dokulara daha kesin tanı koymalarına ve sınıflandırmalarına yardımcı olmayı amaçlamaktadır. Bu matematiksel araçlar, daha tutarlı ve nesnel yapılar sunarak gözlemci-içi ve gözlemciler-arası değişkenliği azaltmaya olanak sağlar. Günümüzde, özellikle dokusal ve/veya yapısal doku analizi temelli otomatik kanser tanı ve sınıflandırması üzerine çok miktarda çalışma bulunmaktadır. Önceki yapısal yaklaşımların farklı tipte dokular için umut verici sonuçlar göstermelerine rağmen, bu yaklaşımlar hücre çekirdeği dışındaki doku bileşenlerinden sağlanabilecek potansiyel bilgiyi kullanabilmekten yoksundurlar. Hâlbuki bu ek bilgi, farklılaşmış bileşenlerden oluşan doku tipleri için ana bilgi kaynaklarından birisini oluşturmaktadır; örneğin lümen bölgeleri, kolon dokusu içindeki bezleri tanımlamaya yardımcı olmaktadır.

Bu tez çalışması, hücre çekirdeği dışındaki doku bileşenlerinin kullanımı için yeni bir yapısal yaklaşımı, yeni bir çeşit kısıtlı Delaunay üçgenlemesini, ortaya koymaktadır. Bu yapısal yaklaşım öncelikle hücre çekirdekleri ve lümen bölgeleri üzerinde iki düğüm kümesi tanımlar. Daha sonra, lümen düğümleri kısıtları oluşturacak şekilde, çekirdek düğümleri üzerinde bir kısıtlı Delaunay üçgenlemesi oluşturur. Son olarak, bu yeni tanımlanan kısıtlı Delaunay üçgenlemesinden çıkarılacak öznitelikleri kullanarak doku örneklerini sınıflandırır.

Elli sekiz farklı hastadan alınan 213 kolon doku örneği üzerinde gerçekleştirdiğimiz deneyler, kısıtlı Delaunay üçgenlemesi yaklaşımı ile eğitim kümesi için yüzde 87.83, test kümesi içinse yüzde 85.71 gibi yüksek doğruluk değerleri elde edildiğini ortaya koymuştur. Ayrıca deneylerimiz, yeni özniteliklerin

tanımlanmasına izin veren bu yeni yapısal gösterimin, kanserli dokuların incelenmesi için daha gürbüz bir çizge-tabanlı yöntem olduğunu ve önceki yöntemlere göre daha yüksek başarı sağladığını göstermektedir.

Anahtar sözcükler: Kısıtlı Delaunay üçgenlemesi, histopatolojik görüntü analizi, otomatik kanser teşhisi ve sınıflandırılması, kolon kanseri, adenokarsinom.

This thesis is dedicated to the memory of my uncle Necati Hançerlioğlu, who devoted himself to the people of his country.

Acknowledgement

"Clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line." (Mandelbrot, 1983).

The subject of science is too broad for one person to be expert in all of its aspects. I thank my master, Assist. Prof. Dr. Çiğdem Gündüz Demir for her special competence in computer sciences and constant support of me throughout this process, and Prof. Dr. Cenk Sökmensüer for his consultancy on medical knowledge. I would also thank all of the members of my thesis committee; Assist. Prof. Dr. Selim Aksoy and Assist. Prof. Dr. Pınar Şenkul for kindly agreeing to be in my thesis committee. I also thank TÜBİTAK-BİDEB and the project under the number TÜBİTAK 106E118 for their financial support to me and our project.

I am extremely grateful to my closest friends for their guidance, encouragement, and just standing by me: İmran Akça, Akın Avcı, Ayça Akçay, Bahadır Bebek, Ceren Bebek, Leyla Bilge, Murat Boğazkesenli, Oğuzhan Çelebi, Pınar Doğan, Hayrettin Gürkök, Burcu Kafa, Bahadır Kemaloğlu, Barış Korkut, Alp Manyas, Çağla Okur, Erkan Okuyan, Yağmur Ökten, Melih Özbekoğlu, Cihan Öztürk, Mustafa Pelit, Akif Burak Tosun, Ömer Sezgin Uğurlu, Hilal Zitouni, and the rest of my friends, who are the most beautiful people in the world.

This thesis, which you are holding in your hands right now, is created on a tough, thorny, everlasting path. The publication of this thesis would be impossible without the endless generosity, patience, care, and love of my parents, Nebahat and Mehmet Erdoğan, my sister Tülay Erdoğan, and my future wife, Bengü Okur. I am deeply indebted to them for just being who they are.

S. Tuncer Erdoğan
July 2009

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer, which is also known as malignant neoplasm, is a serious, lethal class of human diseases that occurs with the ungoverned expansion, division, and spread of abnormal cells. 12.5 percent of deaths worldwide is due to cancer and cancer results in more deaths than AIDS, tuberculosis, and malaria combined, as presented in Table 1.1. It is also the second leading cause of death in economically developed countries, following heart diseases, and the third leading cause of death in developing countries, right after heart diseases and diarrhoeal diseases [52].

Cancer affects a variety of organs or systems. Most types of cancer, such as prostate, breast, and colorectal, form a tumor and affect the organ they originate from. On the other hand, some types do not form a tumor, like leukemia. One of the most common tumor-forming-cancer type is colon cancer, which is also named as colorectal cancer or large bowel cancer. According to studies, it is the third leading cause of cancer-related deaths in developed countries for both men and women [52].

According to World Health Organization, one third of these cancer incidents could be reduced by enforcing cancer-preventing strategies [127]. Another third could be cured if they are diagnosed early and treated adequately. Regular screening examinations by health care professionals may prevent the cancer to be formed and result in the removal of pre-malignancy growths. Considering the occurrence

	Worldwide		Developing		Developed	
	Rank	%	Rank	%	Rank	%
Heart diseases	1	19.6	1	18.1	1	28.6
Malignant neoplasms	2	12.5	3	10.2	2	26.2
CV diseases	3	9.6	4	9.5	3	9.9
Lower resp. infections	4	6.7	5	7	4	4.4
COPD*	5	4.8	8	4.9	5	3.8
HIV/AIDS	6	4.6	6	5.3	-	0.3
Perinatal conditions**	7	4.5	7	5.1	-	0.4
Diarrhoeal diseases	8	3.2	2	16.1	-	0.1
Tuberculosis	9	2.9	9	3.3	-	0.2
Road traffic accidents	10	2.0	-	2.2	9	1.5
Malaria	11	2.1	10	2.5	-	0.0
Diabetes mellitus	12	1.7	-	1.6	7	2.6
Suicide	13	1.6	-	1.5	8	1.6
Cirrhosis of the liver	14	1.4	-	1.4	10	1.5
Measles	15	1.4	-	1.6	-	0.0

The number zero in a cell indicates a non-zero estimate of less than 500 deaths.

*COPD is chronic obstructive pulmonary disease.

**This cause category includes "causes arising in the perinatal period" as defined in the International Classification of Diseases, principally low birthweight, prematurity, birth asphyxia, and birth trauma, and does not include all causes of deaths occurring in the perinatal period.

Source: Lopez AD, Mathers CO, Ezzati M, et al. Global and regional burden of disease and risk factors, 2001: Systematic analysis of population health data. Lancet. 2006;367(9524):1747-57.

Table 1.1: Leading causes of death worldwide (in developing and developed countries), 2001

and death rate of cancer throughout the world, the value of early and accurate detection of the cancerous tissues and the selection of the correct treatment plan are very important.

It should be noted that the correct treatment selection is always the key to recovery and convalescence. One of the main factors that affects the treatment selection is accurate diagnosis and grading of cancer. In the current practice of medicine, several methods have been proposed for cancer diagnosis. The first category of these methods consists of medical imaging techniques, such as magnetic resonance imaging (MRI), nuclear MRI (NMRI), computed tomography (CAT scan), and positron emission tomography (PET scan). These techniques are used to diagnose the cancerous regions, but they are incapable of providing reliable information for grading process. The second group of formerly proposed cancer diagnosis methods is molecular diagnosis [33, 38, 65, 78, 110]. This type of diagnostic methods is not for widespread use, due to the fact that genetic information is highly complicated and there exists a requirement for both specialists and complex and costly apparatus.

For these reasons, in the current practice of medicine, histopathological examination is still the gold standard for both cancer diagnosis and grading. For the examination, a sample tissue, which is called biopsy, is surgically removed from a patient. Afterwards, the biopsy is placed onto a glass slide and stained with a special technique to enhance contrast in the microscopic image. In the histopathological examination, a pathologist examines the structure of the tissue to determine whether it is cancerous or not. If it is cancerous, s/he also determines the type and grade of cancer.

Histopathological examination yields valuable clinical information and provides accuracy both in diagnosis and grading [14, 22, 63, 69]. Neverthless, cancer diagnosis is still a major challenge for cancer specialists worldwide [95, 102]. The main drawback of the histopathological examination is that the analysis is subjective to visual interpretation and experience of the pathologist, especially in grading process [3, 36, 114].

To decrease the subjectivity level, and thus, to help pathologists make more

reliable decisions, computer-aided diagnosis has been proposed. Computer-aided diagnosis is becoming more robust and reliable with the development of novel approaches and evolution of algorithms in the long run. They also gain momentum and become widely accepted with the falling prices and improvement of hardware infrastructure. Cheap processing power is coming into the picture as a natural result of these maturing computer technologies. However, the existing systems present their own challenges and have their own shortcomings.

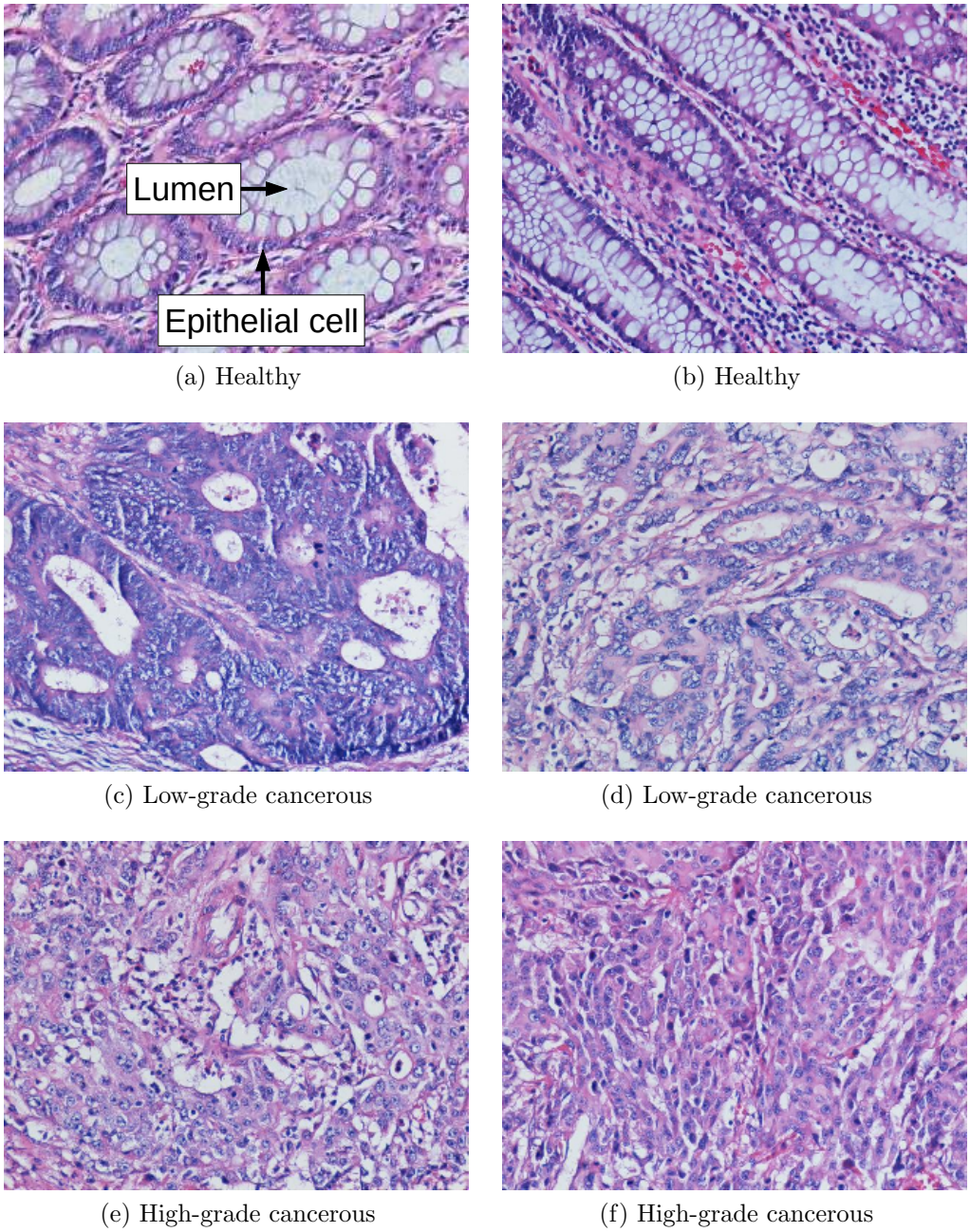
1.1 Motivation

Ongoing development in computer technologies has already been canalized into cancer research projects. Many studies have been proposed to use computerized image analysis to support pathologists and to reduce the variability between the decisions of the pathologists. In these studies, a tissue is represented with a set of mathematical features and these features are then used in automated diagnosis and grading process. These studies mainly focus on textural and/or structural tissue analysis.

In the first group of these studies, the texture of the entire tissue is characterized with a set of textural features such as those calculated from co-occurrence matrices [40, 43, 103], run-length matrices [126], multiwavelet coefficients [67, 126], fractal geometry [7, 37, 44], and optical density [43, 126].

In the second group, the cell distribution within a tissue is represented as a graph and structural features are extracted from this graph representation. Previous studies consider the locations of cell nuclei as nodes to generate such graphs including Delaunay triangulations [40, 72], Gabriel graphs [112, 124], minimum spanning trees [18, 124], and probabilistic graphs [32].

The major drawback of the previous graph-based studies is their incapability of using potential information that is provided by other tissue components rather than cell nuclei. Because of their nature, such information becomes useful especially for the representation of the tissue types where tissues consist of hiearchical



(a) Healthy (b) Healthy

(c) Low-grade cancerous (d) Low-grade cancerous

(e) High-grade cancerous (f) High-grade cancerous

Figure 1.1: Histopathological images of colon tissues. These tissues are stained with the hematoxylin-and-eosin technique, which is routinely used to stain biopsies in hospitals.

structures. For example, in colon tissues, epithelial cells are lined up around a lumen to form a glandular structure. The gland architecture for a normal colon tissue is shown in Figures 1.1a and 1.1b with the lumen and an epithelial cell of a single gland being indicated with arrows. Colon adenocarcinoma, which accounts for 90–95 percent of all colorectal cancer incidents, distorts the gland formations. At the beginning, the degree of distortion is lower such that the gland formations are well to moderately differentiated; examples of such low-grade cancerous tissues are shown in Figures 1.1c and 1.1d. Then the distortion level becomes higher such that the gland formations are only poorly differentiated; examples of such high-grade cancerous tissues are shown in Figures 1.1e and 1.1f. For the automated diagnosis and grading of colon cancer, these distortions should be quantified. Obviously, for a colon tissue, the additional information obtained from luminal components facilitates better tissue quantification compared to the case where the information is obtained from only cellular/nuclear tissue components.

Another drawback of the previous graph-based techniques is their requirement of high-quality segmentation. Accurate extraction of nuclei information is crucial for the proposed algorithms and it surely ensures higher and more reliable recognition rates, because the algorithms make use of spatial information provided by these nodes. However, in the image magnification on which a graph is extracted, the boundaries of individual cell nuclei of colon tissues are occasionally uncertain and they are even inseparable by an expert. Figure 1.2 points out the indistinguishability of some nuclei groups. Therefore, An alternative node definition algorithm is necessary to eliminate the requirement of a high-quality segmentation.

1.2 Contribution

The formation of glandular structures is characterized with the locations of cells and luminal regions with respect to each other and it affects the decisions of pathologists for diagnosis and grading. Thus, it should also affect the results acquired by computer-aided diagnosis systems. However, traditional graph-based

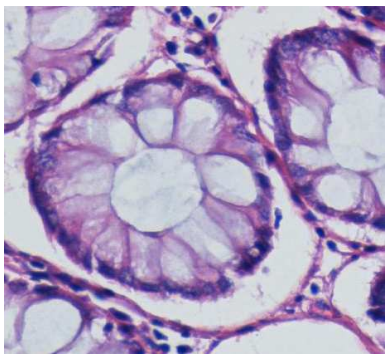


Figure 1.2: In the image magnification on which a graph is extracted, the boundaries of individual cell nuclei of colon tissues are occasionally uncertain.

approaches ignore this fact and use the information provided by the features that are extracted from any type of graph which is just built onto cell nuclei structures. On the contrary, for better characterization of a tissue, it is beneficial to define representative graph nodes on luminal regions rather than defining them only on nuclear regions. These nodes can be used in such a way that they facilitate the luminal information for recognition systems.

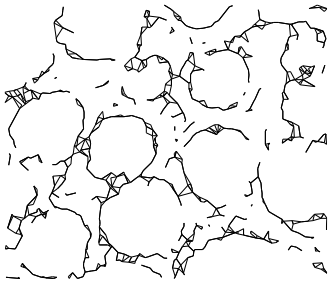
In this thesis, we report a new structural method that considers the locations of both nuclear and luminal components for tissue representation. Unlike the previous studies that use the standard Delaunay triangulation and its corresponding Voronoi diagram on nuclei, we propose to use a new type of constrained Delaunay triangulation (and its corresponding Voronoi diagram) to represent the tissue. In this representation, we assign edges between nuclear components where luminal components form the constraints. Then we define a new set of structural features on this constrained Delaunay triangulation, and use these features in the classification of colon tissues. The constrained Delaunay triangulation of an exemplary colon tissue image, which is shown in Figure 1.3e, and its corresponding Voronoi diagram are shown in Figures 1.3a and 1.3b, respectively. In their construction, both nuclear and luminal components are considered as opposed to the construction of standard Delaunay triangulation and Voronoi diagram where only nuclei components are utilized but non-nuclei components are not considered. For the same tissue image, the standard Delaunay triangulation and its

corresponding Voronoi diagram are shown in Figures 1.3c and 1.3d, respectively.

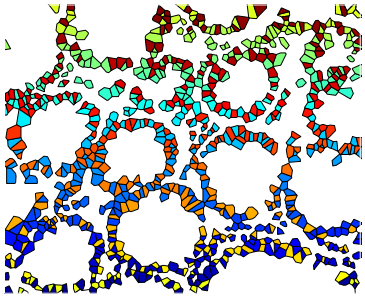
In this representation, a set of circular primitives, formed by a technique called circle-fit transform [115], is used as the nodes of the triangulation. With the help of this approach, the problem that arises from the necessity of using classical and inefficient segmentation algorithms is alleviated. Furthermore, images with lower resolution are sufficient for the proposed circle-fit transform, which decreases the CPU-time.

1.3 Organization of the Thesis

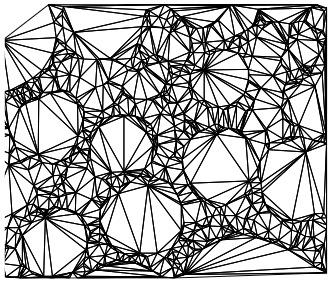
The remaining of this thesis is organized as follows: In the first section of the following chapter, a brief explanation of medical background and terminology is presented. The following section exposes the previous studies in this research area of cancer diagnosis, emphasizing their disadvantages for colon tissues, and revises the related work on the use of constrained Delaunay triangulation. Chapter 3 explains the details of our proposed structural method. Consequently, Chapter 4 describes the experimental framework and analyzes the experimental results. Finally, Chapter 5 provides a summary of our work and discusses the future directions of our study.



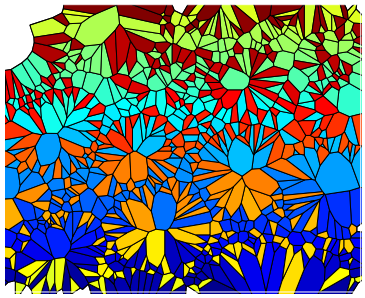
(a) Constrained Delaunay triangulation



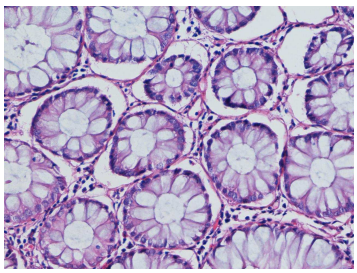
(b) Voronoi diagram of constrained Delaunay triangulation



(c) Delaunay triangulation



(d) Voronoi diagram of Delaunay triangulation



(e) Tissue image

Figure 1.3: A sample Delaunay triangulation built onto nuclei structures of colon tissues. It can be observed that nuclei around or in the middle of a luminal region are connected to each other in the traditional Delaunay triangulation, where they are not suppossed to be.

Chapter 2

Background

In this chapter, the underlying medical terminology of this thesis is presented. The structure of colon tissues, staining process, structural changes in different grades of colon cancer are some of the topics investigated in the first section.

Following the medical background, a brief overview of the previous computational methods for the diagnosis of cancerous tissues other than the graph-based methods is given. Afterwards, broader knowledge on the subject of graph-based methods is presented. Finally, other applications of constrained Delaunay triangulation are overviewed.

2.1 Medical Terminology

2.1.1 Colon tissues

The *colon*, or the large intestine or large bowel, is the last portion of the digestive track and is responsible for extracting water and electrolytes from feces just before they are excreted. The remaining solid waste is also stored until leaving the body through the anus.

When cancer is suspected, a variety of methods can be applied to detect the status of a tissue. In the current practice of medicine, *histopathological examination* is the gold standard. For diagnostic evaluation, a sample tissue is surgically removed from the patient. This procedure is called a *biopsy*. The *biopsy specimen* is sent to a *pathologist* for the microscopic examination.

In the next step, sections are taken from the biopsy specimen and they are stained with special chemical compounds for the ease of visibility at microscopic level. In this thesis, we use the images of colon tissues stained with the *hematoxylin-and-eosin technique* (H&E). This staining technique is a conventional one and is routinely used to stain tissues at hospitals. In this technique, hematoxylin is the active ingredient of the staining solution and colors nucleic acids with a blue-purple hue. On the other hand, alcohol-based acidic eosin colors eosinophilic structures (i.e., proteins) with bright pink. Consequent view under the microscope carries the characteristic blue-stained nuclei and pink-stained stroma [47, 75]. Therefore the color spectra of the images of tissues stained with the H&E technique are commonly rich of blue-purple, pink, and white pixels.

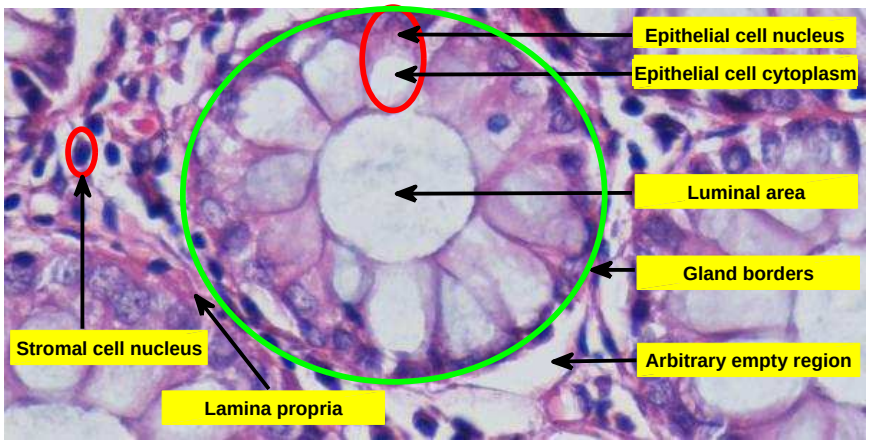


Figure 2.1: The layout of a colon tissue stained with the H&E technique

Figure 2.1 exhibits the layout of a colon tissue stained with the H&E technique. Cells in a tissue could be grouped into two: *epithelial cells* and *stromal cells*. Simple columnar, non-ciliated *epithelial cells* form the epithelium tissue of

colon [96]. A sample epithelial cell is highlighted in Figure 2.1 with a red circle, of which the dark purple region constitutes its *nucleus*, and the lighter white section next to nucleus is *cytoplasm*. A group of epithelial cells is lined up around an oval vacant region called *luminal area* to form a *gland* all together. Luminal area of a gland body is also marked in this figure. Also, there may exist an arbitrary vacant region around or inbetween gland bodies, which is just an artifact that arises from the sectioning procedure.

The remaining region outside the gland bodies consists of loose connective tissue, which makes up the support structure of biological tissues and holds all of the structures in the tissue together. The cells found in the loose connective tissue could be called as stromal cells. These are diffused around glandular bodies and they are not part of a gland structure. Moreover, the pink area around stromal cells is composed of noncellular material called *lamina propria*.

2.1.2 Colon adenocarcinoma

Colon tissues may easily become cancerous in course of time, due to the fact that an uninterrupted stress provided by storaged solid waste is present. The risk factors of colon cancer also include a diet low in fiber and high in fat, certain types of colorectal polyps, inflammatory bowel disease such as Crohn's disease or ulcerative colitis, smoking, alcohol, and some inherited genetic disorders transmitted at birth. Improper nutrition habits tending to pervade in developing countries, as a natural result of intense life routines, heighten the risk of colon cancer. Colon adenocarcinoma is its most common type, which accounts for approximately $90-95$ percent of all colon cancers.

In a typical healthy colon tissue, there exists an harmonious and coherent composition of gland bodies and loose connective tissue. With the abnormal growth and division of cells, due to the aforementioned reasons, *tumors* may grow out of healthy tissues and consistent structure of colon may get corrupted. A tumor can be *benign* (non-cancerous) or *malignant* (cancerous). If the tumor is malignant and has expansionary properties, it may diffuse into other organs and

become lethal in case of not being cured. Tumor *grading* is the scheme used to catalogue the status of cancer cells in the sense of how anomalistic they appear and how rapidly the tumor is likely to grow and spread.

Tumors may be graded on four-tier, three-tier and two-tier scales [46]. In two-tier grading scheme, which is the scheme used in this thesis, there are

- *Low-grade cancer*, in which the glands are well to moderately differentiated, and
- *High-grade cancer*, in which the glands are only poorly differentiated.

2.2 Previous Studies on Tissue Analysis

Many studies have been proposed to use computerized image analysis to support the pathologists and to reduce the variability between the decisions of pathologists. Previous studies have focused on different aspects of image processing. Mainly, these studies made use of textural and structural representation of histopathological images. Less number of studies used other types of information such as color based features and/or morphological features. In this chapter, we will overview the fundamental textural and structural features of histopathological image processing.

2.2.1 Textural features

In the first group of previous studies, the texture of the entire tissue is characterized with a set of textural features such as those calculated by the following approaches:

Co-occurrence matrix : Co-occurrence matrix is accumulation of pixel level data which is the distribution of co-occurring values at varying offsets. It is initially defined by Haralick et al. in 1973 [61]. The other known identities are

co-occurrence distribution, gray-level co-occurrence matrix (GLCM), and spatial dependence matrix. For a $w \times h$ image I parameterized by an offset $(\Delta x, \Delta y)$, the co-occurrence matrix C is defined in Equation 2.1:

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

Co-occurrence matrix is sensitive to rotation, so the use of a set of offsets corresponding to $0, 45, 90$ and 135 degrees results in a degree of rotational invariance. Also, the value of the image is generally referred to the gray-scale value of the specified pixel. However, in literature, co-occurrence matrices have been commonly used to extract texture information not only from gray-scale images [8, 18, 42, 43, 60, 103, 124, 126], but also from colored images [107, 108].

The resulting co-occurrence matrices are not sufficient to be analyzed by themselves, so many features are extracted from these matrices such as inertia, entropy, total energy, angular second moment, contrast, correlation, variance, sum average-variance-entropy, and difference variance-entropy-moment, also known as Haralick's features [61].

Run-length matrix : Gray-level run-length method has been initially defined by Galloway [51] and its features are then extended by Chu et al. [19]. It is another effective way of accessing higher order statistical texture features [2]. Although it is shown that run-length method is slightly less adequate for texture analysis comparing to other methods [21, 122], later improvements consolidate the employability of the algorithm [113].

Galloway's proposition of run-length matrix is as follows: For a given image, a run-length matrix $p(i,j)$ is defined as the number of runs with pixels of gray-level i and run length j, where a run is defined as a set of linearly sequential pixels belonging to the same intensity. In literature, there exist some features defined on run-length matrices such as short runs emphasis, long runs emphasis, gray-level non-uniformity, run length non-uniformity, run percentage, low gray-level run

emphasis, and high gray level run emphasis [113]. Later, additional features like short run low gray-level emphasis, short run high gray-level emphasis, long run low gray-level emphasis, and long run high gray-level emphasis are also defined on these matrices [26]. High gray-level run emphasis (HGRE) feature is presented as an example in Equation 2.2:

$$HGRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p\,(i,j) \cdot i^2 \tag{2.2}$$

These features are also used for the diagnosis and grading of cancerous tissues in previous studies conducted by Weyn et al. [125] and Bibbo et al. [8]. Extracted features are used for classification with KNN classification [126], histogram method [125], or three-way discriminant analysis [8].

Multiwavelet coefficients : Wavelets are special classes of functions that are used to represent data or other functions by dividing them into different scale components [54]. They have very important applications especially in image (data) compression and denoising (image enhancement) [5, 15, 24, 25, 27, 55, 66, 83, 87, 90, 93, 99]. The details of the wavelets are out of the scope of this thesis and references will be left to the reader for future investigation.

Three classes in which wavelet transforms are divided are continuous, discrete, and multiresolution-based wavelets. Multiwavelets are superior to former classes and possess additional properties like orthogonality, symmetry, and vanishing moments, which are known to be important in signal and image processing, and resulting in being advantageous over scalar ones [1, 111, 116].

Multiwavelet coefficients enable researchers to lower the dimension of data and analyze it through some features. Some set of features like energy and entropy can be extracted from multiwavelet coefficients and they are also used to improve accuracy on colon image classification in previous studies [29, 67, 123, 126].

Fractal dimension : Mandelbrot defined a fractal as an image and/or geometric shape that appears identical and repeats itself as it is scaled down, which

cannot be represented by classical geometry [80]. In fractal geometry, the fractal dimension, D, is a measure of complexity which gives an indication of how a fractal appears to fill space as it is zoomed further. The statistical quantity D is acquired with the following equation, where $N(\epsilon)$ is the number of self-similar structures of diameter ϵ necessary to cover the structure:

$$D = \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log \frac{1}{\epsilon}} \tag{2.3}$$

Previous studies on histopathological image analysis make use of fractal dimension information to improve accuracy achieved by other methods [7, 37, 44]. Although the analysis shows that fractal dimension is highly correlated with features like correlation and entropy, it is shown that fractal dimension information improves sensitivity and can be useful for automated techniques in clinical practice in the future [44].

2.2.2 Structural features

In the second group of previous studies, the cell distribution within a tissue is represented as a graph or diagram and structural features are extracted from these representations. These studies are mainly focused on the following representations:

Voronoi diagram : Voronoi diagram is defined as the decomposition of a regular plane with n data points into convex polygons such that each polygon encloses its unique origin point and every point inside is closer to the origin point of its polygon than any other data points. For each point p_i in the set of coplanar points P, a boundary enclosing all the intermediate points lying closer to the corresponding origin point p_i than other points in the set P can be drawn. Each one of the enclosing points are called Voronoi polygons. Figure 2.2 demonstrates the schema of a Voronoi diagram constructed on 50 random points.

The very first Voronoi-like diagram was proposed by René Descartes when he

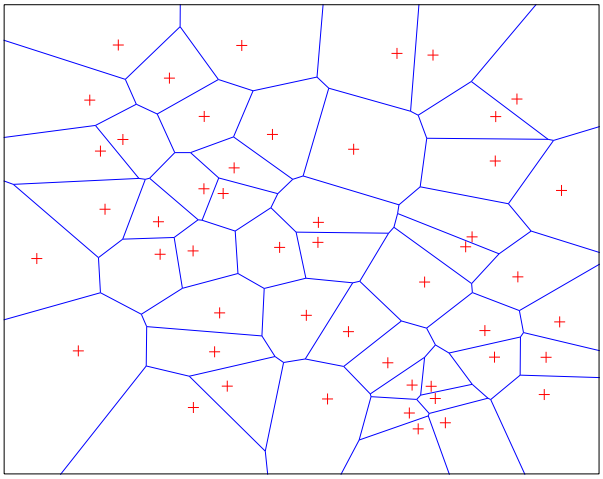


Figure 2.2: A Voronoi diagram of random points. It is observable that every single point has its own mutually disjoint convex polygon, i.e., Voronoi polygon.

was trying express the division of universe by stars in 1644 [88]. The original definition of Voronoi diagrams goes back to 1850 with Lejeune Dirichlet [35], but later, Voronoi extended the investigation of Voronoi diagrams to higher dimensions in 1907 [120]. Due to its history, Voronoi diagrams are also called Dirichlet tessellation (also medial axis transform, Wigner-Seitz zones, domains of action, and Thiessen polygons [98]). Each mutually disjoint convex polygon of a Voronoi diagram is called Dirichlet regions, Thiessen polytopes, or Voronoi polygons.

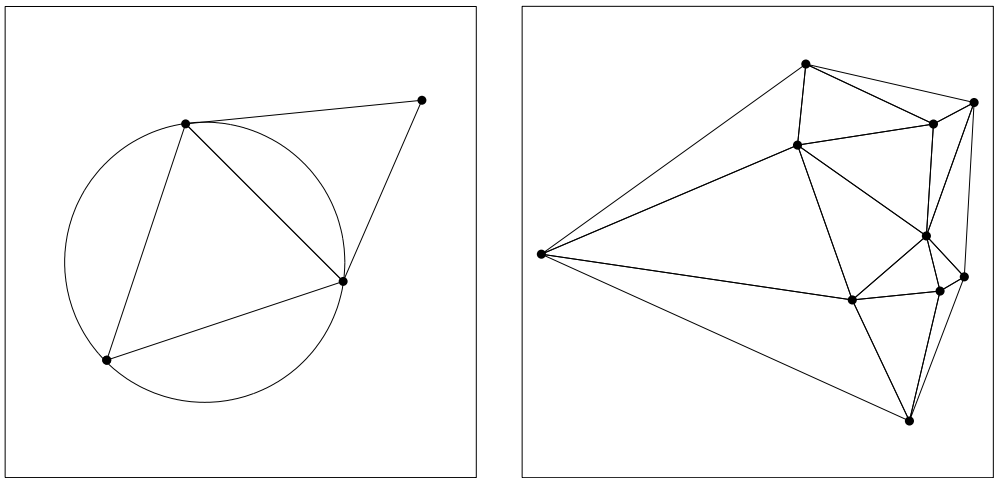
Voronoi diagram has become a classical approach and has been applied across many studies in computer vision and image analysis [6, 45, 88, 98]. Intrinsically, cancer studies have taken advantage of the Voronoi diagrams [9, 72, 81, 101, 124, 126]. In these studies, the very fundamental features such as area and shape of the polygons are sufficed to ensure the satisfying discriminative power among other types.

Delaunay triangulation : Delaunay triangulation is a special type of graph of which edges satisfies the following rule: The circumcircle of each individual triangle is an empty circle, pointing out that there must not exist any other point inside the circle. The triangulation is named after Boris Delaunay, who defined

it in 1934 [30]. The original formal definition of a single Delaunay triangle for two-dimensional space is as follows:

Definition 1 *Let* $\mathbf{P}$ *be a finite set of points. Non-collinear points* $\mathbf{p_i}, \mathbf{p_j}$ *and* $\mathbf{p_k}$ *of set* $\mathbf{P}$ *form a Delaunay triangle* $\mathbf{t}$ *if and only if there exists a location* $\mathbf{x}$ *which is equally close to* $\mathbf{p_i}, \mathbf{p_j}$ *and* $\mathbf{p_k}$ *and closer to* $\mathbf{p_i}, \mathbf{p_j}, \mathbf{p_k}$ *than any other* $\mathbf{p_m} \in \mathbf{P}$*. The location* $\mathbf{x}$ *is the center of a circle which passes through the points* $\mathbf{p_i}, \mathbf{p_j}, \mathbf{p_k}$ *and contains no other points* $\mathbf{p_m}$ *of* $\mathbf{P}$*. For the 2D space, there exist only one circle which is the circumcircle of* $\mathbf{t}$ *[48].*

Figure 2.3a visualizes the circumcircle of a single Delaunay triangle. Delaunay triangulation is the dual of Voronoi diagram in $\mathbb{R}^2$. It maximizes the minimum interior angle of all of the angles of the triangles in the triangulation. Delaunay triangulation of ten random points is visualized in Figure 2.3b, and presented together with its corresponding Voronoi diagram in Figure 2.4:



(a) Circumcircle of a Delaunay triangle

(b) Delaunay triangulation of ten random points

Figure 2.3: Delaunay triangulation

There are many features available to be calculated using Delaunay triangulation, since it is a variation of graphs. Most of the features below are also common to the following graph representations, which are to be explained in this section. For each feature type, some statistics such as mean, standard deviation, skewness,

and kurtosis can be calculated over the corresponding feature values, creating a broad selection pool of subfeature types.

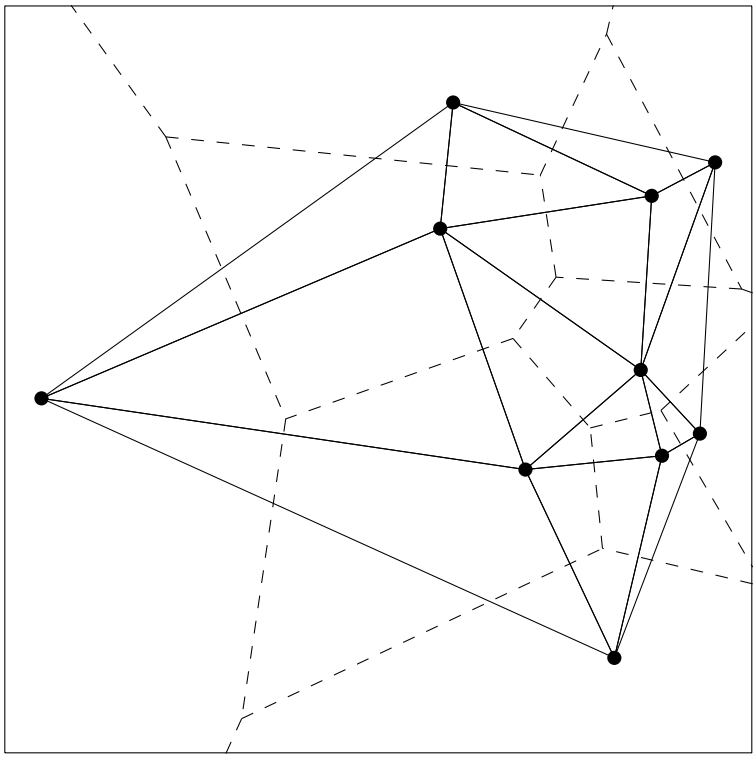


Figure 2.4: Delaunay triangulation together with its corresponding Voronoi diagram. In this figure, dashed lines represent the boundary lines for Voronoi polygons, and the solid lines belong to Delaunay triangulation.

- Length of edges - Features can be defined on the information of distance between the nodes which share a common edge. The feature can be defined on global scale (e.g., the mean of edge lengths), or defined per node (e.g., the standard deviation of edge lengths for each node). An alternative feature can be the distance to nearest neighbors per node (which is connected by an edge) or the distance to the farthermost neighbor [72, 124]. Another modification of the feature can be the average distance to, for instance, the five closest neighbors [71].

- Number of edges - Similar to the previous feature, total number of edges in the entire image or a variation of statistical data per node can be used for classification [72, 124].

- Triangle area - The mean of the triangle area for the entire image or the kurtosis of areas of triangles that share a common vertex per node are two examples of how this feature can be utilized [72].
- Number of triangles - The number of triangles per unit area or per node can be other types of features [72].
- Polarity of the edges - Angles inbetween the edges can also be utilized for polarity information [18].

Not only Delaunay triangulation, but all of the graph types used in the previous studies are built on a set of nodes which, in fact, represents the positional information of nuclei of the tissues. Previous studies have adopted this approach as a general rule and defined features on these graphs [10, 18, 40, 72, 124]. As mentioned before, most of the aforementioned features are also common to these graphs.

Gabriel graphs : Gabriel graph is another type of neighborhood graph which attempts to represent the overall spatial arrangement of the points in a set of P, named after Gabriel who proposed the use of newborn graph in 1969 [50]. It is a subgraph of Delaunay triangulation [94]. Gabriel graph forbids inclusion of any other point to the circle accepting an edge between two points as diameter, unlike the Delaunay triangulation, which defines the circle on three points forming a Delaunay triangle. A Gabriel graph edge is defined as follows:

Definition 2 *Let* $\mathbf{P}$ *be a finite set of points. Two points* $\mathbf{p_i}$ *and* $\mathbf{p_j}$ *of set* $\mathbf{P}$ *are connected by an edge of the Gabriel graph, and they are said to be Gabriel neighbors, if and only if the circle having line segment* $\mathbf{p_i p_j}$ *as its diameter does not contain any other points of* $\mathbf{P}$ *in its interior.* $\mathbf{DG}(\mathbf{P})$ *contains the Gabriel graph of* $\mathbf{P}$, $\mathbf{GG}(\mathbf{P})$.

A sample Gabriel graph of the point set presented in Figure 2.5a, is given in Figure 2.5c. It is obvious that this Gabriel graph is a subgraph of Delaunay triangulation presented in Figure 2.5b.

Gabriel graphs have been used in various applications, but have not been employed widely in cancer diagnosis applications. Sudbo et al. [112] and Weyn et al. [124] made use of Gabriel graphs to increase accuracy in their work.

Minimum spanning tree : A spanning tree T of connected, undirected, and unweighted graph G is a subgraph that reaches out to all the nodes with $n-1$ edges and is a tree, which is a connected graph without cycles. On the other hand, a minimum spanning tree of a weighted graph is the one of the alternative spanning trees with the least total weight. A minimum spanning tree for the given point set is also constructed and presented in Figure 2.5d.

Similar to the Gabriel graphs, the minimum spanning tree are used to improve discriminative power of image analysis systems in the previous studies of Choi et al. [18] and Weyn et al. [124].

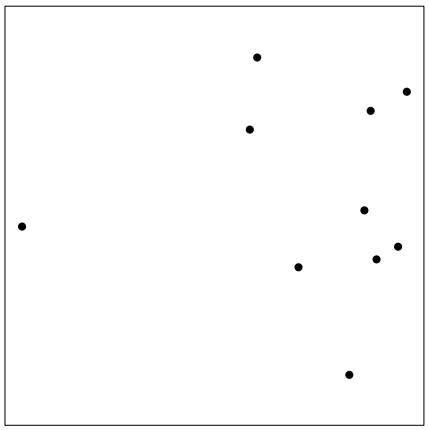
Probabilistic graphs : Unlike the previous graph types, probabilistic graphs are constructed by probabilisticly assigning edges between every pair of nodes. Then the cell distribution in a tissue is quantized by the features extracted from these probabilistic graphs [32, 59]. For example, in [32], the probability of the existence of an edge $P(u,v)$ is defined as:

$$P(u,v) = d(u,v)^{-\alpha} \tag{2.4}$$

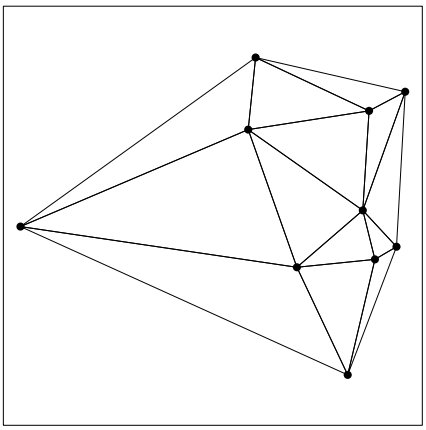
where $d(u,v)$ is the Euclidean distance from node u to node v, and $0 < \alpha \leq 1$ is the parameter that controls the edge density of the graph. Larger values of α increase the number of edges in the graph. Formal definition of these probabilistic graphs is as follows; a sample of a probabilistic graph is given in Figure 2.5f.

Definition 3 *Let $G = (V, E)$ be a generated graph with V being the set of nodes and E representing the edges of the graph. The binary relation E of V is defined as $E = \left\{(u,v) : r < d(u,v)^{-\alpha}, \forall u, v \in V\right\}$, where r is a generated random real number between 0 and 1.*

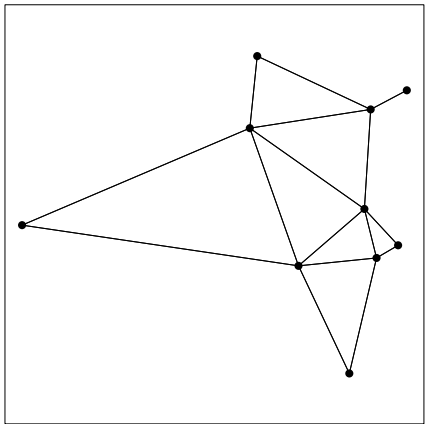
Augmented (complete) graphs : Augmented graphs are undirected,



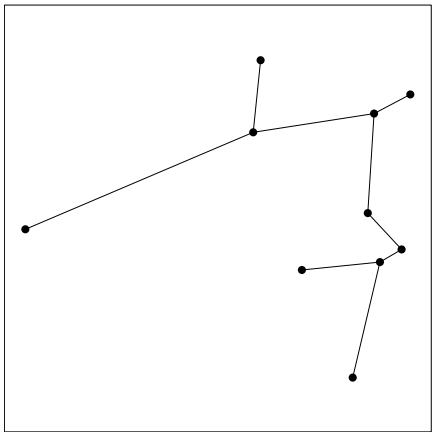
(a) Point set



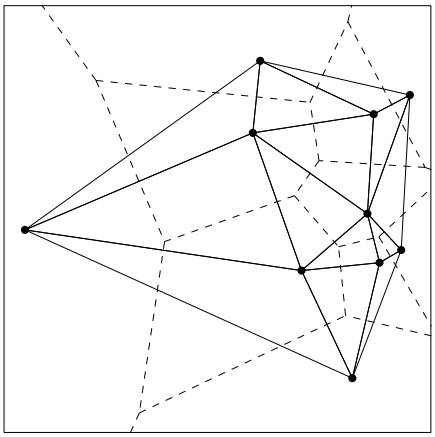
(b) Delaunay triangulation



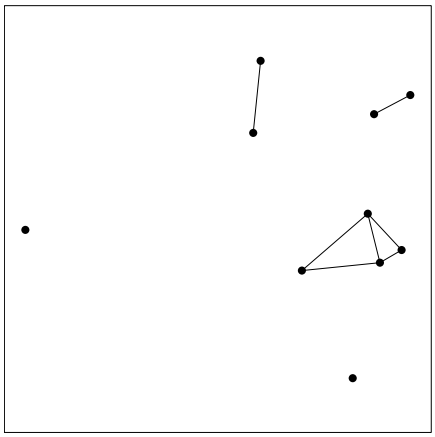
(c) Gabriel graph



(d) Minimum spanning tree



(e) Voronoi diagram with Delaunay triangulation



(f) Probabilistic graph with $r = 0.1$

Figure 2.5: Various structural types of computational geometry

weighted, complete graphs without self loops. With the use of augmented graphs in a study by Demir et al. [31], usage of several control parameters are eliminated due to the fact that augmented graphs are complete graphs. Inclusion of all possible edges between every pair of nodes prevents the loss of any kind of existing spatial information. The study revealed a remarkable and exceptional accuracy achieved in the classification of glioma, which is a type of brain cancer.

2.3 Constrained Delaunay Triangulations

As its definition is given earlier in Section 2.2.2, Delaunay triangulations are one of the most commonly used structural entities in the history of mathematics and computational geometry. There exist an extension to Delaunay triangulation, constrained Delaunay triangulation, which satisfies the following properties:

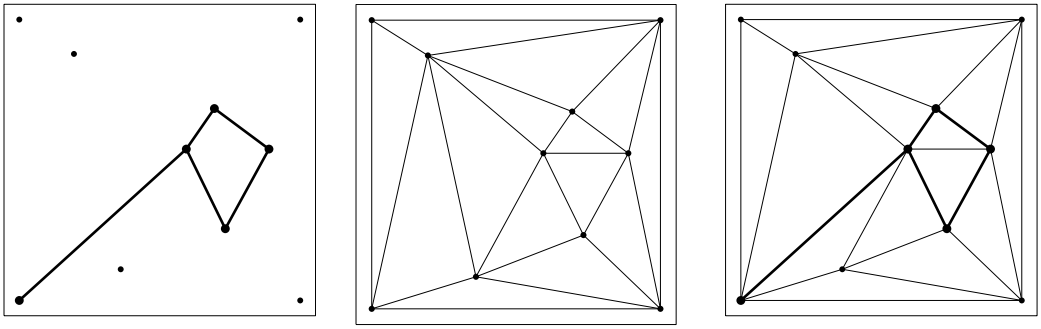
- The prespecified edges that are obliged to be included in the final graph representation appear in the triangulation.
- The triangulation is as close as possible to the Delaunay triangulation.

More formally, constrained Delaunay triangulation is originally defined by Chew in [17] as follows:

Definition 4 *Let G be a straight-line planar graph. A triangulation T is a constrained Delaunay triangulation (CDT) of G if each edge of the G is an edge of T and for each remaining edge e of T, there exists a circle C with the following properties: (1) the endpoints of edge e are on the boundary of C, and (2) if any vertex v of G is in the interior of C then it cannot be "seen" from at least one of the endpoints of e (i.e., if you draw the line segments from v to each endpoint of e then at least one of the line segments crosses an edge of G).*

Figure 2.6a demonstrates ten points in $2D$ space. Suppose that the thick lines are the constraint edges and they must appear in the final triangulation. Figure

2.6b shows the conventional Delaunay triangulation calculated with empty-circle rule, disregarding the constraints. On the other hand, Figure 2.6b points out the structuring of constrained Delaunay triangulation. Note that the prespecified edges are included in the final triangulation, and the graph is kept as close as to Delaunay triangulation.



(a) Data points and the constraint edges

(b) Delaunay triangulation of data points

(c) Constrained Delaunay triangulation of data points and constraint edges

Figure 2.6: Constrained Delaunay triangulation

Constrained Delaunay triangulation is introduced by Lee et al. [74] and Chew [16] separately, but later the definition has been expanded to higher dimensions by other studies [56, 105, 109]. There have been many studies about the construction of Delaunay triangulation from scratch and reconstruction of deformed or swelled triangulations. Construction of constrained Delaunay triangulation can be achieved via adaptation of proposed methods. Delaunay construction algorithms and current succesful adaptations are categorized under three general groups:

- Divide and conquer algorithms: In divide and conquer algorithms, the input set is partitioned into subsets, the smaller subsets of the input set are triangulated individually and the resulting triangulations are merged in the final step. These type of algorithms are generally sophisticated and require complex data types, but they are well suited for parallel programming. Chew [16], Dwyer [41], Ruppert [97], Hardwick [62], and Cingoni et al. [20] developed algorithms for triangulation in the context of divide and

conquer approach. The runtime of these algorithms is generally $O(n \log n)$, buth Dwyer proposed an algorithm with $O(n \log \log n)$ expected time on uniformly distributed sites.

- Sweep-line algorithms: Sweep-line algorithms (or plane sweep algorithms) use an imaginary line (generally a vertical line for the sake of simplicity) that is sweeped across the set of nodes. One of the regions that the sweep-line separates is being processed as the line moves further. Every time the sweep line encounters a point $p_m \in P$ or the status changes due to the change of a considered criterion, the triangulation is enhanced to cover the new point.

  First sweep-line algorithm for Voronoi diagram and Delaunay triangulation was proposed by Fortune in 1987 [49], with a complexity of $O(n \log n)$. Later, Shewchuk [106], Domiter and Žalik [39, 130] have improved the sweep-line algorithms for constrained Delaunay triangulation.

- Incremental algorithms : This category holds incremental insertion and incremental search algorithms. Incremental algorithms have probably been the simplest and most popular algorithms for constructing the Delaunay triangulations. In the construction phase, new vertices or edges are inserted iteratively, and in every iteration, the empty circle rule is obeyed. Incremental update of a complex, previously constructed constrained Delaunay triangulation with addition of new nodes or deletion of nodes is also feasible in these algorithms. Guibas launched the studies in incremental algoritms of Delaunay triangulations with the studies [58] and [57] in 1985 and 1992, respectively. Later, quite a few studies came on the scene and developed the school of incremental algorithms [4, 34, 70, 73, 76, 117, 119, 118, 129, 131].

- Other triangulation algorithms include high-dimensional embedding, convex-hull based algorithms, and gift-wrapping algorithms [130], sparse matrix algorithms, quadtree, and Dewall.

Constrained Delaunay triangulations (CDTs) have played an important role in practical applications of diverse fields, including:

- Surface modelling and 3D object reconstruction: Delaunay triangulation (DT) is famous for producing high quality triangular mesh, in which the triangles are comparably neat and elongated triangles are eliminated. Thus, surface modelling and 3D object reconstruction have been important areas in which DTs and CDTs deployed widely. Floriani and Puppo have developed a new algorithm for multiresolution surface description [28]. Moreover, triangulated irregular networks (TIN) make use of the benefits of the Delaunay triangulation and reflect the surface morphology of terrain in various applications in geographical information systems (GIS) [39, 117]. Xue et al. matured the idea of reconstruction of three-dimensional complex objects for geological research [128]. Other studies around the topic include the study by Park et al., who developed a system based on incremental CDT algorithms to compress TIN data [89]. Also, Muckell et al. developed the idea of using CDT together with a modified TIN to create a hydrology-aware triangulation of terrain data [86].

- Finite element methods: Shenoy et al. made use of CDTs to construct a finite element mesh with the representative atoms as nodes. In [104], the linear triangular finite elements are utilized to link atomistic and continuum models. In another paper, Lu and Dai developed another approach for the construction of CDTs for the sake of meeting demands of multichip module layout design [77].

- Segmentation: Hu et al. demonstrated the success of a novel method using image foreground and background estimation based on the CDT. Background seed estimation, and noise suppression are independently developed on top of a CDT algorithm [64]. Although the study is dependent on the success of face and torso detection, an interesting methodology in the use of CDTs has been exhibited.

- Motion planning: Motion planning problems involve the dynamic nature of constraints, because in robotics, for example, collusion-free path may change in time due to the movement of obstacles. Kallman et al. developed a fully dynamic CDT algorithm, which allows the degenerations and repairs itself automatically in case of edge overlapping or self-intersections [70]. In

the study, other applications of the dynamic CDT algorithm in visualization, geometric modelling, reconstruction, GIS are also emphasized.

Other areas that applications of CDTs can be found include image processing systems such as skeletonisation [85] or network routing [100]. It must be taken into account that limitless numbers of studies in Delaunay triangulations are available while the new fields of application of CDTs are being discovered recently.

Chapter 3

Methodology

In Chapter 1, we have mentioned that histopathological examination is subjective to visual interpretation and experience of a pathologist. To decrease the subjectivity level, and thus, to help pathologists make more reliable decisions, computer-aided diagnosis has been proposed. There have been many methods and studies on the subject of computer-aided diagnosis, given in Chapter 2. However, graph based studies in this area are still lack of utilizing other components rather than cell nuclei in favor of obtaining higher classification accuracy. In this chapter, we introduce a new type of constrained Delaunay triangulation for the purpose of fulfilling the expectations of a gap-sensitive methodology, which in our case lumen entities in a colon tissue.

The proposed method comprises a series of processing steps. In the first step, the pixels of the image are clustered into three groups which correspond to nuclei, cytoplasm, and white (including lumen) areas using the k-means clustering algorithm. After a slight preprocessing is applied, the following step fits circular objects into these enclosed spaces of white areas and nuclei pixels with the help of circle-fit transform. The outcome constitutes the input, white and purple nodes, to the next step. Afterwards, a Delaunay triangulation is built on the set of nodes initially. A graph is constructed out of the initial triangulation with the use of our novel constraint definition on constrained Delaunay triangulation, and various features are extracted from this resulting graph. Finally, training

and classification of the images is accomplished by using these feature sets. The overall system architecture is shown in Figure 3.1.

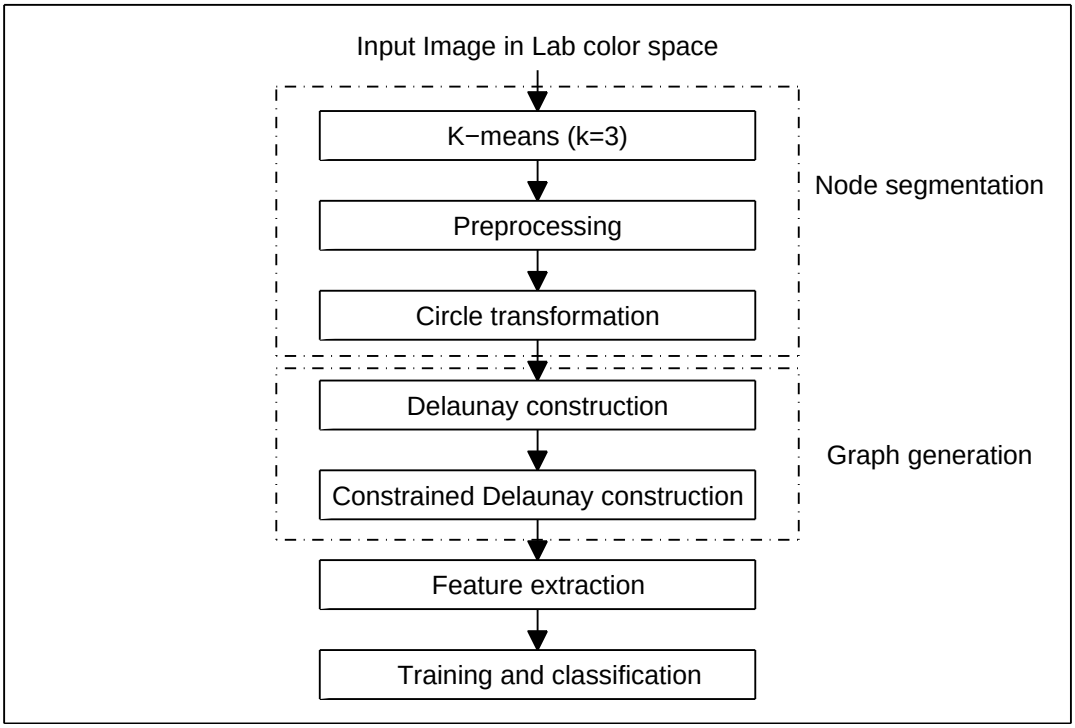
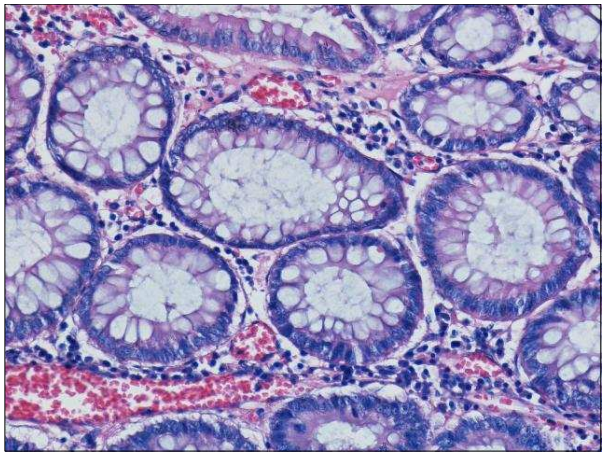


Figure 3.1: Overall system architecture

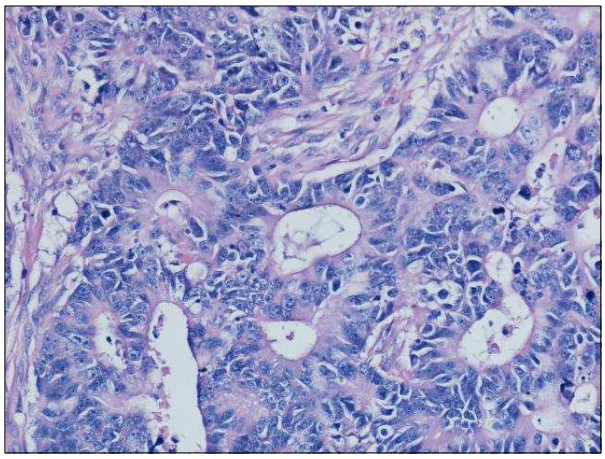
As demonstrated in Figure 3.1, the system architecture consists of four main components: Node segmentation, graph generation, feature extraction, and classification. Details of these components are given in the rest of this chapter. Throughout these sections, biopsy images shown in Figures 3.2a, 3.2b, and 3.2c will be used to demonstrate the output of the corresponding step. They are the images of the tissues that are healthy, low-grade cancerous, and high-grade cancerous, respectively.

3.1 Node Segmentation

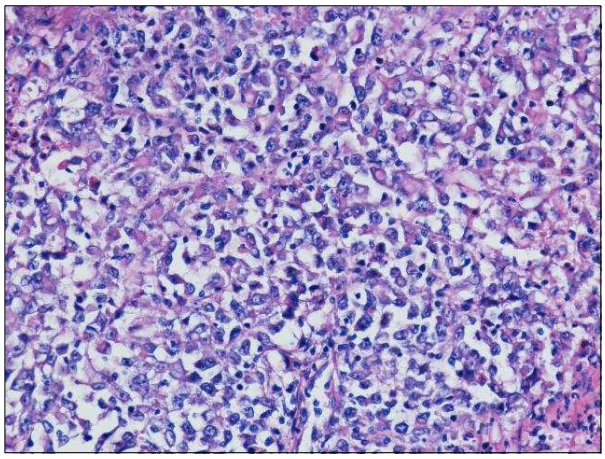
Increasing the accuracy of diagnosis and grading in graph-based approaches is directly concerned with proper positioning of nodes. High-quality node segmentation results in a more representative graph which reflects the architecture of



(a) Healthy



(b) Low-grade cancerous



(c) High-grade cancerous

Figure 3.2: Colon tissue samples

the tissue it is built on better. It also reduces the effects of artifacts, noise, and color variation due to hematoxylin-and-eosin (H&E) staining. In our problem, for colon tissues stained with the H&E staining technique, it may sometimes be impossible to cope with indistinguishability of nuclei from each other. Besides, we also need symbolic white nodes which will represent the lumen structures. For the node segmentation problem, we follow the steps whose details are given below.

3.1.1 Transforming into Lab color space

Lab color space is a color-opponent space, which is originally aforethought to estimate cognitive vision of homogeneity of human. The color space, with the components $\boldsymbol{L}$ for lightness, $\boldsymbol{a}$ and $\boldsymbol{b}$ for the color-opponent dimensions, improves the computerized image analysis methods with its $\boldsymbol{L}$ component closely matching human perception of lightness [53].

The segmentation component of our system first transforms the RGB biopsy images to Lab color space for the histopathological analysis. The intent was to simulate human perception of color at first, and later experiments proved the advantages of the use of Lab color space over RGB color space, and more accurate segmentation is provided by this color space.

3.1.2 K-means clustering

The well known k-means clustering algorithm is a process to partition N-dimensional population into k sets and it has been one of the simplest unsupervised learning algorithms [79]. The algorithm attempts to locate cluster centers of k sets by minimizing the sum of distances over all clusters, where different distance approaches are proposed. Basically, the sum of squared Euclidean distances from points to cluster centroids is accepted as the distance function in our system. With k being the desired number of clusters; $S_i, i = 1, 2, \ldots, k$ being the clusters; and μ_i being the centroid of all points $x_j \in S_i$, Equation 3.1 presents

the distance function:

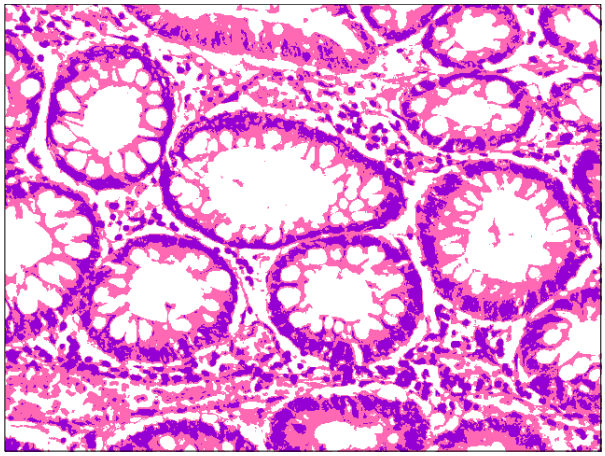
$$D = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \tag{3.1}$$

The k-means algorithm is useful to discriminate the white, nuclei and cytoplasmic regions' pixels of the biopsy images. In our problem, the k-means algorithm is applied on the images with $k = 3$ to differentiate these three disjoint regions. Right after detecting which pixels belong to which regions, average L values of the regions are used to determine the type of cluster vectors (white, nuclei or cytoplasm). In Lab color space, $\boldsymbol{L}$ represents the lightness of color, which yields 0 for black and 100 for diffuse white. Hence, the cluster vector with the highest L value, and its corresponding pixels can be labeled as white and the darkest one and its corresponding pixels can be labeled as nucleus, leaving the remaining one left and its pixels as cytoplasm, which has a pink color in RGB space.

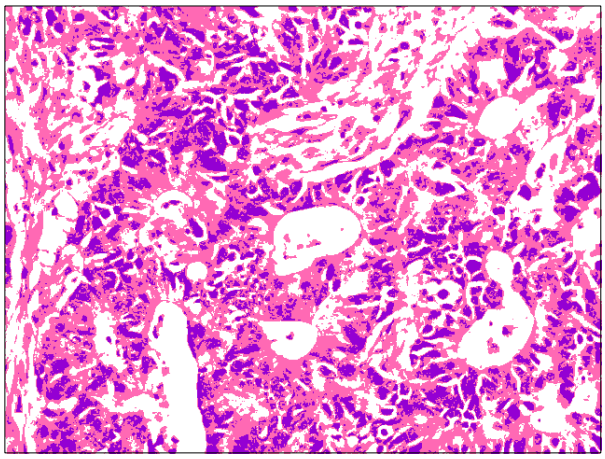
It is observed that selecting k value as 3 is adequate and sufficient for our problem, because the H&E staining technique dyes chromatin-rich nuclei regions with dark purple, eosinophilic cytoplasm regions (belonging to stromal cells and connective tissues) with pink, and releases the vacant regions as it is, white. Consequently, the k-means algorithm easily recognizes and distinguishes pixels of three dissimilar zones in the image, and Lab conversion also increases the rate of separation. The clustered output of the k-means algorithm for the sample tissue images (Figure 3.2) are presented in Figure 3.3.

3.1.3 Preprocessing

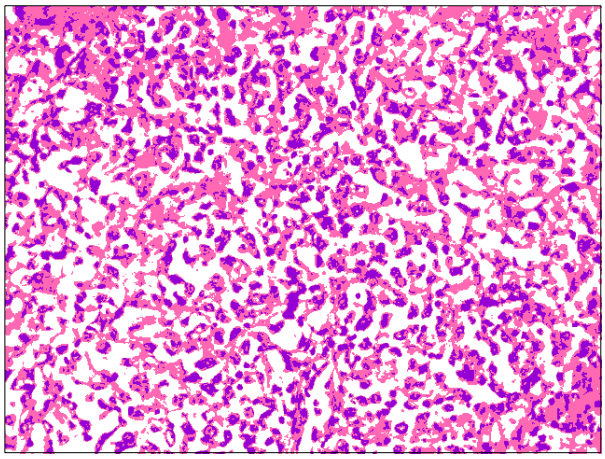
After k-means clustering, we have labeled maps (pixels), which have noise on a small scale. To eliminate noise, we have preprocessed these maps and applied some morphological operations. These operations include a morphological close operation followed by a morphological open operation with a flat, square structuring element with edge length of 3 (for the histopathological images at the resolution 640×480). Close operation removes small gaps and eliminates noisy



(a) Healthy



(b) Low-grade cancerous



(c) High-grade cancerous

Figure 3.3: Clustered biopsy samples

data, while open operation disqualifies the undesired results of close operation, turning the regions back into their original penetration and scope.

3.1.4 Circle-fit transform

With the execution of the k-means clustering algorithm and preprocessing step, the pixels of the original image are classified and appointed to the correct pixel group. Although the pixel classification provides the essential information about tissue distribution, the outcome itself does not contribute to nuclei and lumen segmentation. The reason is that, given earlier in Section 1.1, the boundaries of individual nuclei of colon tissues are occasionally uncertain and they are even inseparable to the experienced specialists' eye. An alternative segmentation and node definition algorithm, which better suits the needs of segmentation of colon tissues, should be applied to eliminate the necessity of a high-quality segmentation.

In literature, there exist some studies which split the segmentation problem into subproblems covering the nucleus segmentation, luminal region segmentation and epithelial cytoplasm segmentation, and then unify the solutions altogether to create a unique, segmented tissue image [82, 91, 92]. However, this approach put forwards the problematic nature of segmentation. They also require high-magnification biopsy images or high-dimensional hyperspectral data. Besides, according to our present knowledge, there do not exist any studies on cytoplasm and lumen segmentation.

From another perspective, the solution to the node segmentation problem might turn into decomposing the clustered image into a series of connected components and using each individual connected component as a separate node. However, our experiments have shown that this methodology results in a significant decrease in classification quality, because of the aforementioned reasons.

To overcome all these problems, we have decided not to segment each actual individual histological component, but approximately represent them. For this

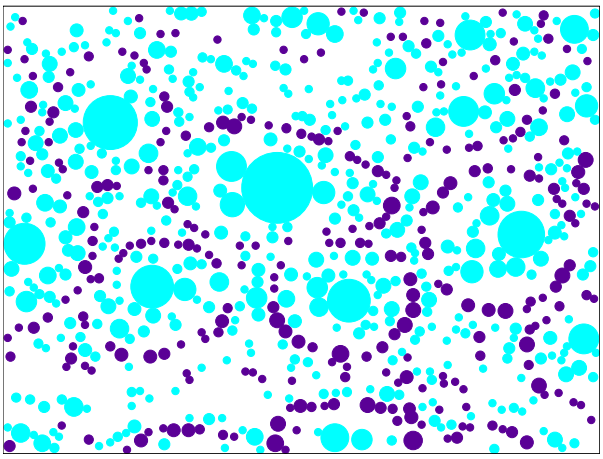
purpose, we have made use of a technique called *circle-fit transform*, which fits the largest circles available into the given connected components. The algorithm of circle-fit transform is detailed in [71, 115].

The motivation behind the use of circle-fit transform is that cell nuclei generally have globular forms, and appear circular in two dimensional microscope scanning. These entities are represented by a single circle in the resulting map of circle-fit transform. On the other hand, wide lumen areas are represented by a single large circle, and/or a few smaller circles inside the luminal area. The circle-fit algorithm removes the circles with a radius below a certain predefined threshold, therefore, this scheme gives rise to the elimination of the artifacts. Epithelial cytoplasms are usually transformed into medium-sized circles around lumen circles, however, the rest of our methodology does not require this distinction. Henceforth, the circles fitted into nuclei regions are referred to as *purple circles*, and those which are fitted into luminal or arbitrary vacant regions around or inbetween gland bodies or epithelial cell cytoplasms are referred to as *white circles*. Fitted circles for the previous sample tissues in Figure 3.2 are presented in Figure 3.4, where cyan circles corresponds to the white circles.

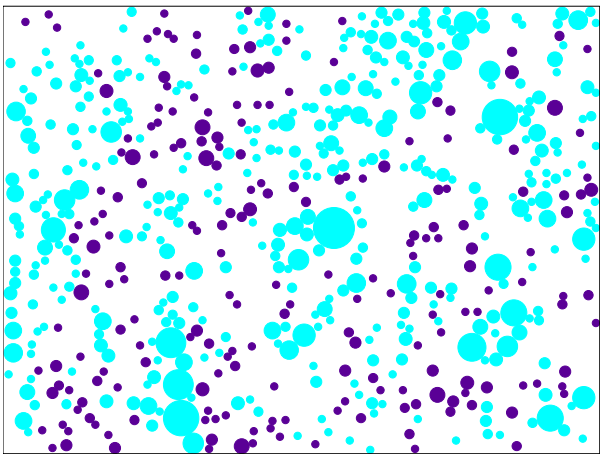
Having two distinct types of nodes, *purple circles* and *white circles*, we may utilize these circles as nodes to enhance the graph generation algorithms. In the next section, we will be discussing the details of our graph generation approach.

3.2 Graph Generation

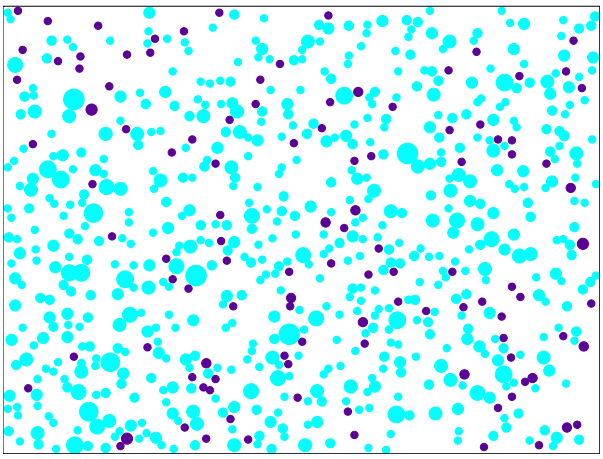
The center of the fitted circles release the coordinates of nodes in two dimensional space, on which we will build our novel graph, constrained Delaunay triangulation. In this section, we will first define the conventional Delaunay triangulation on the set of nodes which reside at the middle of white circles and purple circles, *white nodes* and *purple nodes*, respectively. Afterwards, we will explain the details of the mechanism that vitalizes the constrained Delaunay triangulation.



(a) Healthy



(b) Low-grade cancerous



(c) High-grade cancerous

Figure 3.4: Circle-fit transform

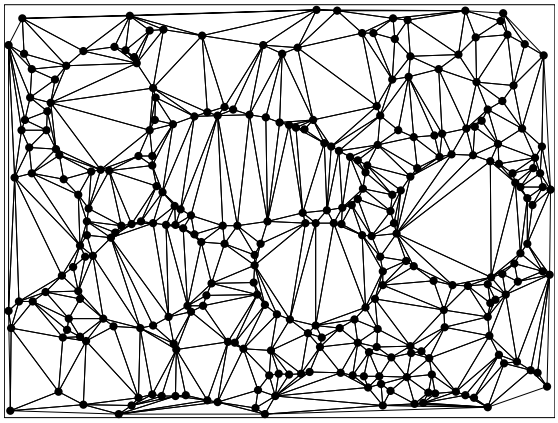
3.2.1 Delaunay triangulation

In previous studies (see Section 2.2.2), Delaunay triangulation algorithm is extensively used for the characterization and description of the tissue. Those studies used Delaunay triangulation, which is built on a set of nodes that represent nuclei of cells in the tissue for the favor of histopathological analysis. The reason behind is that Delaunay triangulation carries the required characteristics which reflects the structuring of the tissue it is built on better than some of the other graph types. Another reason is that the generality of the Delaunay triangulation shall provide the fundamental skeleton to the other types of graph. For example, a minimum spanning tree of a set of points P is a subgraph of Delaunay triangulation of P.

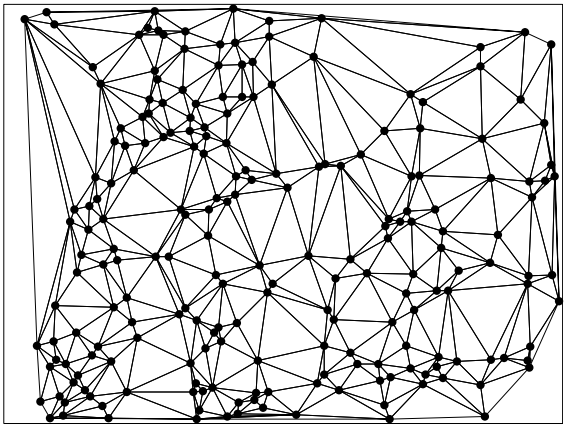
These former approaches have been beneficial for tissues without dominant hierarchical structures and have provided a distinctive set of features that increases the diagnosis and grading accuracy. However, not every tissue is plain as those like liver or lymph node tissues. For the tissues with hierarchical structures such as colon tissues with gland structures, these approaches are insufficient in recognizing such structuring of the tissue on its entirety. Presented in Figure 3.5, the outcome is not that representative enough to analyze the status of glandular structures.

As a solution to this problem, one may think of utilizing the representative nodes we have defined on the other tissue components, as explained in the previous section. For this purpose, we first form a Delaunay triangulation on the set of nodes composed of purple nodes and white nodes. The formal definition of this Delaunay triangulation is given as follows:

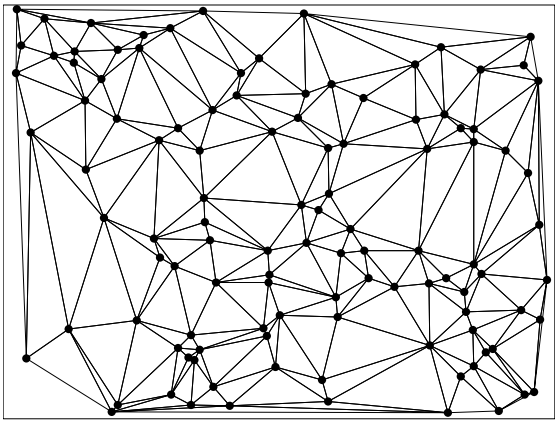
Definition 5 *Let* $\mathbf{S_p}$ *be the finite set of primary nodes and* $\mathbf{S_s}$ *be the finite set of secondary nodes. Non-collinear points* $\mathbf{p_i}, \mathbf{p_j}$ *and* $\mathbf{p_k}$ *of set* $\mathbf{S_w} \cup \mathbf{S_p}$ *form a Delaunay triangle* $\mathbf{t}$ *if and only if there exists a location* $\mathbf{x}$ *which is equally close to* $\mathbf{p_i}, \mathbf{p_j}$ *and* $\mathbf{p_k}$ *and closer to* $\mathbf{p_i}, \mathbf{p_j}, \mathbf{p_k}$ *than any other* $\mathbf{p_m} \in \mathbf{S_w} \cup \mathbf{S_p}$. *The location* $\mathbf{x}$ *is the center of a circle which passes through the points* $\mathbf{p_i}, \mathbf{p_j}, \mathbf{p_k}$ *and contains no other points* $\mathbf{p_m}$ *of* $\mathbf{P}$. *For the* $2D$ *space, there exists only one circle, which is*



(a) Healthy



(b) Low-grade cancerous



(c) High-grade cancerous

Figure 3.5: Delaunay triangulation constructed on only the purple nodes

the circumcircle of $\mathbf{t}$.

Figure 3.6 shows the resulting Delaunay triangulation of the set of nodes that are given in Figure 3.4. In this figure, white and purple nodes are represented with empty circles and dots, respectively.

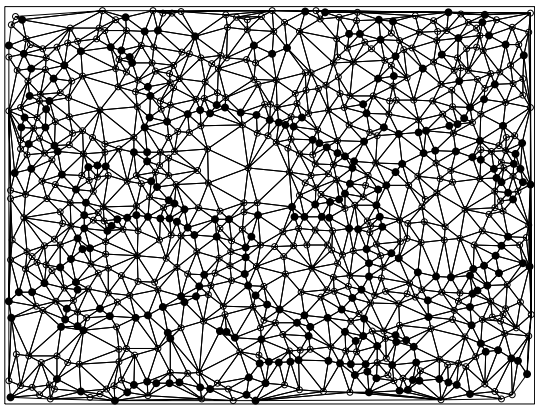
3.2.2 Constrained Delaunay triangulation

The advantages of the utilization of white nodes are explained in the previous section. However, any features that can be obtained from the resulting Delaunay triangulation are no more significant than those that was obtained from the former definitions of Delaunay triangulation. For example, the average edge length in a graph, where the representative white nodes and purple nodes are connected, does not provide better understanding of the health status of the tissue.

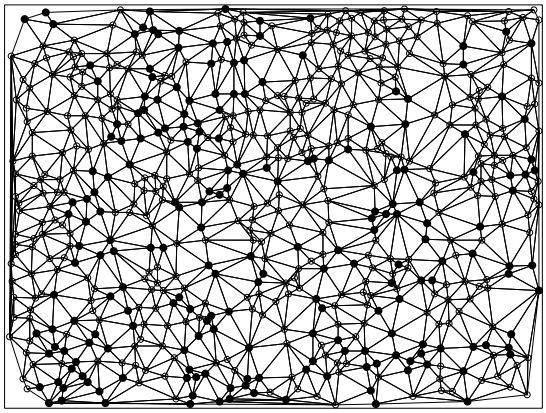
To this end, we have defined novel constraints on the constrained Delaunay triangulation and used that graph representation for cancer diagnosis and grading of colon tissues. This triangulation has the following properties:

- The triangulation includes both white and purple nodes for its initial construction.
- The initial triangulation treats the white and purple nodes in the same way, that is the difference of these nodes is not considered and the joint set of these two types of nodes is processed for the construction.
- After the initial triangulation, the white nodes and the edges that are connected to at least one of the white nodes are deleted from the triangulation.

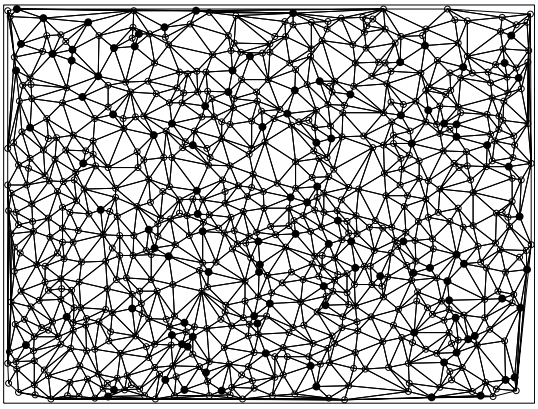
More formally, the constrained Delaunay triangulation can be defined as follows:



(a) Healthy



(b) Low-grade cancerous



(c) High-grade cancerous

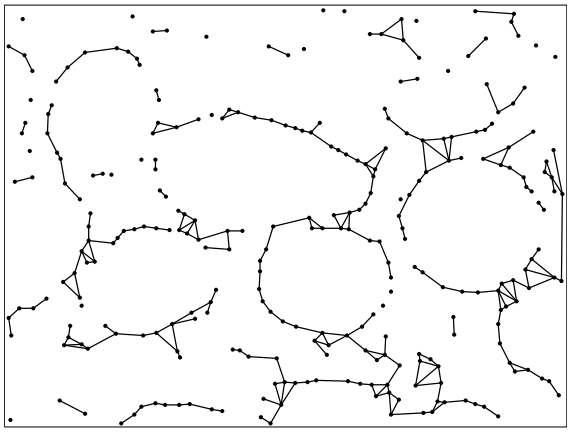
Figure 3.6: Delaunay triangulation constructed on both the white and purple nodes

Definition 6 *Let* $\mathbf{S_p}$ *be the finite set of primary nodes,* $\mathbf{S_s}$ *be the finite set of secondary nodes, and* $\mathbf{E_c}$ *be the final edges of the constrained Delaunay triangulation. Non-collinear points* $\mathbf{p_i}, \mathbf{p_j}$ *and* $\mathbf{p_k}$ *of set* $\mathbf{S_w} \cup \mathbf{S_p}$ *has the right to form a Delaunay triangle* $\mathbf{t}$ *if and only if there exists a location* $\mathbf{x}$ *which is equally close to* $\mathbf{p_i}, \mathbf{p_j}$ *and* $\mathbf{p_k}$ *and closer to* $\mathbf{p_i}, \mathbf{p_j}, \mathbf{p_k}$ *than any other* $\mathbf{p_m} \in \mathbf{S_w} \cup \mathbf{S_p}$*. The location* $\mathbf{x}$ *is the center of a circle which passes through the points* $\mathbf{p_i}, \mathbf{p_j}, \mathbf{p_k}$ *and contains no other points* $\mathbf{p_m}$ *of* $\mathbf{P}$*. For the* $2D$ *space, there exists only one circle which is the circumcircle of* $\mathbf{t}$*. An edge* $\mathbf{e_{ij}}$ *is an element of* $\mathbf{E_c}$ *if and only if it also is an edge of any of the aforementioned triangles and* $\mathbf{p_i}, \mathbf{p_j} \in \mathbf{S_p}$*.*

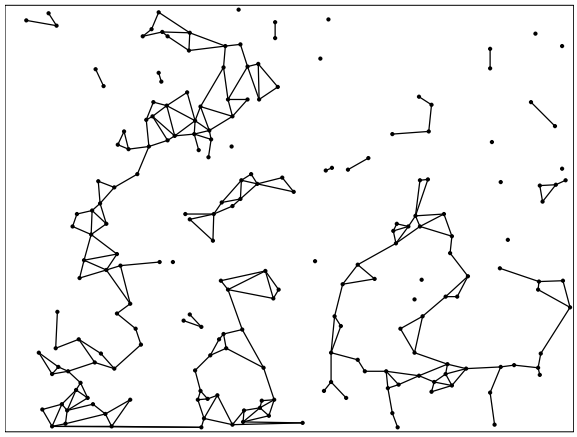
To make it clear, the final triangulation for the aforementioned healthy, low-grade cancerous, and high-grade cancerous tissues is given in Figure 3.7. This figure exposes the differentiation of triangulation characteristics with the development of cancer. The constraints preserve the prolonged tracks of gland borders for the healthy tissues (Figure 3.7a), while the triangulation reveals the fact of gland thickening for the low-grade cancerous tissues (Figure 3.7b). As the cancer develops and the distortion of the glandular structures increases, the number of constrained edges decreases (in fact, the number decreases down to zero and edges vanish in some samples), due to the fact that individual cells become more and more isolated from each other (Figure 3.7c).

3.3 Feature Extraction

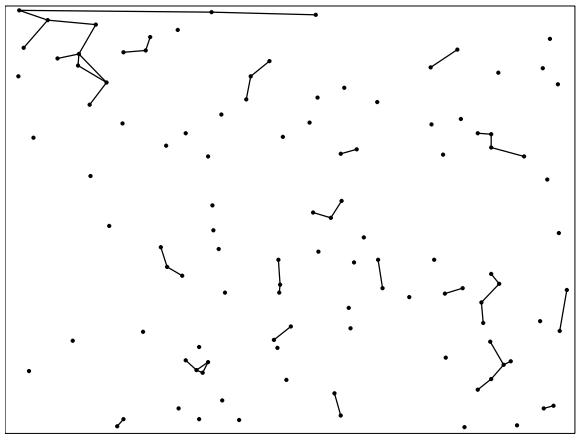
After the construction of constrained Delaunay triangulation, we have a set of primary (purple) nodes S_p and a set of edges E_c. We shall define a set of features on this triangulation for the classification of histopathological images. Feature extraction step makes both use of the classical Delaunay triangulation and constrained Delaunay triangulation and extracts many quantitative features, which represent the qualitative measures considered by pathologists.



(a) Healthy



(b) Low-grade cancerous



(c) High-grade cancerous

Figure 3.7: Constrained Delaunay triangulation

3.3.1 Connectivity based features

Figure 3.2 shows the deviations in the tissue with the development of cancer and it points out that cancer damages the association among cells. The conformity of cells decays with the distortion level, and cells belonging to glands become to get isolated. Hence, connectivity characteristics of the nodes can be used for the understanding of distortion in gland structures. The following set of extracted features reflects the structural properties of the tissue, and thus, provides information about its organizational characteristics. For this purpose, we first define the terms and expressions and give the definition of the features:

The neighborhood N_i for a node p_i is defined as its immediately connected neighbors:

$$N_i = \{p_j : e_{ij} \in E_c\} \tag{3.2}$$

For each individual node $p_i \in S_p$ in a triangulation, *node degree* $d(p_i) = |N_i|$ is the number of nodes that node p_i is connected with an edge $e_{ij} \in E_c$.

In an undirected and unweighted graph, the distance between two nodes is the number of edges in any one of the shortest paths connecting these two nodes. This is also known as the *geodesic distance*, because it is the length of the graph geodesic between those two nodes [13, 84]. Besides, within graph theory and network analysis, there are various measures of the centrality of a node within a graph, which is based on the idea of geodesic distance, that determine the relative importance of a node within the graph. The definitions of the extracted features are given as follows.

- Average degree : Averaging the node degrees on the whole triangulation presents a good indicator of the tissue connectivity. Given the degree of a node $d(p_i)$, average degree in a triangulation is

$$\frac{\sum_{p_i \in S_p} d\left(p_i\right)}{|S_p|} \tag{3.3}$$

- Average degree for nodes with $d\left(p_i\right) \geq 2$: In the graph representation of the tissue, there exist some *isolated nodes* with a degree of $d\left(p_i\right) = 0$. These nodes generally correspond to isolated cells in luminal areas in healthy samples, shown in Figure 3.8, or the poorly-differentiated colon carcinoma cells. Additionally, there exist some *end nodes* with $d\left(p_i\right) = 1$, which constitute the boundaries of the clustered cells. Including the degree of these cells in the calculation of average degree may affect the quality of the feature, so we chose to separate these nodes from the calculation. For this purpose, we have defined another feature that excludes these nodes, and computes the average degree of nodes with a degree of at least 2.

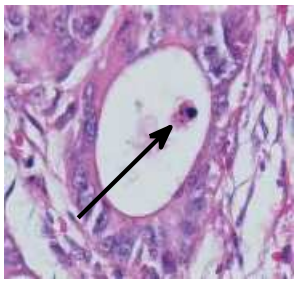


Figure 3.8: An isolated node in a luminal area

- Isolated node number and end node number : As we have decided to separate the isolated nodes and end nodes from the rest for the purification of information, we also include this information to the feature set. To this end, the number of isolated nodes with $d\left(p_i\right) = 0$, and the number of end nodes with $d\left(p_i\right) = 1$ in a triangulation is also included in the recent feature set.

- Average clustering coefficient for nodes with $d\left(p_i\right) \geq 2$: Clustering coefficient is a metric to compute how a node and its neighbors tend to

become a complete graph. The metric, introduced by Watts and Strogatz in [121], is defined as follows:

Definition 7 *The clustering coefficient C_i for a node p_i is the percentage of the existing edges between the neighboring nodes divided by the number of edges that could potentially exist between these nodes. For an undirected graph, if a node p_i has $|N_i|$ neighbors, $\frac{|N_i|\cdot(|N_i|-1)}{2}$ edges could exist among the nodes within its neighborhood. Thus, the clustering coefficient C_i is*

$$C_i = \frac{2\,|\{e_{jk}\}|}{|N_i| \cdot (|N_i| - 1)} : V_j, V_k \in N_i, e_{jk} \in E \tag{3.4}$$

The clustering coefficient estimates the cliquishness of a typical neighborhood of a node, hence it exactly is a local property. With the use of the definition above, we have defined a global property, the *average clustering coefficent for nodes with $d(p_i) \geq 2$*, as the arithmetic mean of clustering coefficients for all but isolated and end nodes in the triangulation.

- Average eccentricity : In graph theory, the eccentricity of a node is defined as follows:

Definition 8 *The eccentricity $\epsilon(p_i)$ of a node p_i in a connected graph G is the maximum geodesic distance between p_i and any other node p_j of G. Given the node $p_i \in S_p$, the eccentricity of the node is:*

$$\epsilon(p_i) = max\{d_g(p_i, p_j) \,|p_j \in S_p\} \tag{3.5}$$

where $d_g(p_i, p_j)$ is the geodesic distance between nodes p_i and p_j.

In short, eccentricity is a measure of the deviation from a common center. In our work, average eccentricity of all nodes is thought to be reflecting the status of global connectivity or the isolation. For nodes with $|N_i| = 0$, the eccentricity value is assumed to be zero.

- Diameter : Given the definition of eccentricity, diameter of a graph is the maximum eccentricity of any node in the triangulation; that is, it is the greatest distance between any two nodes. Diameter $D(S_p)$ is defined as:

$$D(S_p) = max\{\epsilon(p_i) | p_i \in S_p\} \tag{3.6}$$

  For a graph with no accessible nodes, the diameter is set to be zero.

3.3.2 Component related features

In a graph, a connected component is a subgraph in which any of its two nodes p_i and p_j are connected to each other with at least one path and no any other node $p_k \in S_p$ can be added to this subgraph. We can extract the following features from this definition:

- Number of components : The number of components in a triangulation might present a good indication of whether the cells are grouped together and form gland structures (as in healthy tissues) or become isolated (as in cancerous tissues).
- Number of components with $n \geq 2$: As we mentioned earlier in average degree feature, we also chose to separate the components with a single node from the rest, and defined this feature as the number of connected components with at least 2 nodes. The number of remaining components with a single node is equal to isolated node number.
- Giant component ratio : Giant component ratio is the proportion of the number of nodes in the connected component with the highest number of nodes to the number of primary nodes in the triangulation.

3.3.3 Spatial features

The spatial properties of the connected nodes can be used for the definition of some features.

- Average edge length and standard deviation of edge length: Average edge length is based on the Euclidean distances between nodes where the location of nodes are the centers of the corresponding circles. The average edge length is calculated as

$$ael(Ec) = \frac{\sum d_E(p_i, p_j)}{|S_p|} : p_i, p_j \in S_p \tag{3.7}$$

  where $d_E(p_i, p_j)$ is the Euclidean distance between nodes p_i and p_j. On the other hand, the standard deviation of edge length is simply the standard deviation of edge length values and given as

$$sdel(Ec) = \sqrt{\frac{\sum \left(d_E(p_i, p_j) - ael\left(E_c\right)\right)^2}{|S_p| - 1}} : p_i, p_j \in S_p \tag{3.8}$$

- Average triangle area and standard deviation of triangle area: In the triangulation, every three nodes forming a triangle gives a good indication of connectivity and positioning of these nodes. The average and standard deviation of these values are also included in the feature set.

In Table 3.1, the summary of the extracted features that we use in our work is given.

3.4 Classification and Feature Reduction

Classification is the last step of our proposed system. So far, we have developed our system to enable ourselves to define biopsy images quantitatively, by extracting distinctive features. After calculating these features, we can train a suitable

Feature type	Feature
*	Average degree
	Average degree for nodes with $d(p_i) \geq 2$
	Isolated node number
	End node number
	Average clustering coefficient for nodes with $d(p_i) \geq 2$
	Average eccentricity
	Diameter
**	Number of components
	Number of components with $n \geq 2$
	Giant component ratio
***	Average edge length
	Standard deviation of edge length
	Average triangle area
	Standard deviation of triangle area
* Connectivity based features ** Component related features *** Spatial features	

Table 3.1: The list of extracted features

classifier and classify new samples according to its quantifiable measures. In this section, we will explain the classifier we select and discuss how we seek for the new ways to develop our system further, including feature reduction.

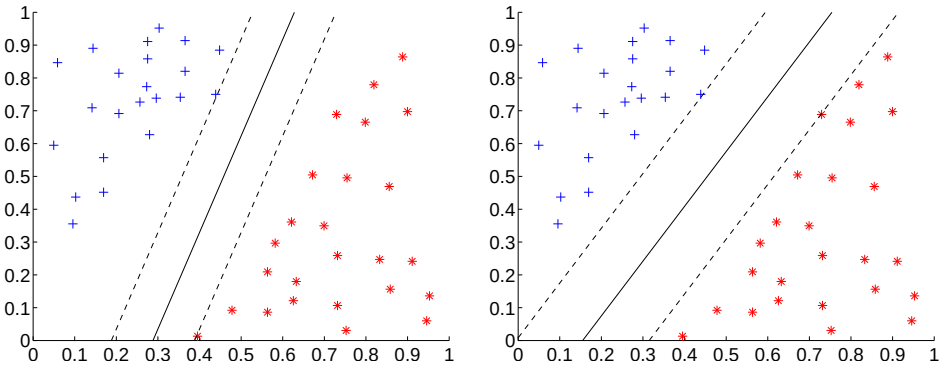
3.4.1 Classification

The definition of features and the selection of a classifier are two important factors that affect the classification accuracy. After the selection of features, we have many options of a classifier to use. We have experimented on some of these classifiers for the classification of colon tissue samples. Among these algorithms, we have chosen the support vector machines (SVM), which perform the best.

3.4.1.1 Support vector machines

Support vector machine (SVM) is a supervised learning algorithm and it is used for regression and classification [23]. SVM basically builds a linear decision surface which separates two sets of data points in an N-dimensional space and it aims to build the one with the maximized margin between these sets. Original proposal of SVM was a linear classifier, which means it is sufficient to build a $(N-1)$-dimensional hyperplane to separate the data. However, with an extension to this original definition, SVM also became a non-linear classifier [12].

Figures 3.9a and 3.9b illustrate the selection of two different hyperplanes (shown with solid lines), which separate two data sets from each other in 2-dimensional space. In these figures, the data points closest to dashed lines are called *support vectors*. The distance between these dashed lines is called the *margin*. The goal of the SVM classifier is to split the space with the most suitable hyperplane, which maximizes the margin. This perfection of division provides a better and more accurate framework when it comes to the classification of unseen data. In Figures 3.9a and 3.9b, both hyperplanes are valid in terms of separating data, but the second one offers a larger margin.

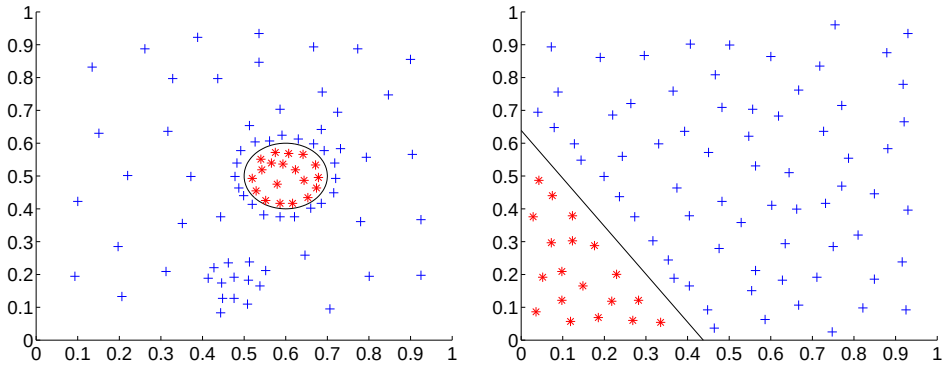


(a) Smaller Margin (b) Larger Margin

Figure 3.9: The hyperplane separating two data sets in 2-dimensional space.

It must be noted that some data sets may not be linearly separable. Figure 3.10a presents a sample of such data set, which has to be separated by a nonlinear region. For such data sets, a *kernel function* is used to map them into a different

space, in which the data points are linearly separable. Figure 3.10b demonstrates the layout of the data points after the transformation. There exists a limitless number of kernel functions, but there are some commonly used ones such as radial basis function (RBF), sigmoid function, and polynomial kernel function.



(a) Original space (b) Feature space

Figure 3.10: Separable classification with kernel mapping.

Dealing with SVM (or any other types of classifier) may bring some problems into equation. To begin with, standard SVM implementation can only deal with binary problems with instances of two classes. We used *Multiclass SVM* in our implementation, because our problem contains three distinct classes: healthy, low-grade cancerous, and high-grade cancerous.

Another problem is that it is not always possible or feasible to provide a perfect split between data sets. If the classifier trains itself beyond its limits to handle every individual sample, it may not generalize well to test samples. The classification accuracy may decrease dramatically, and this effect is called *over fitting*. To avoid over fitting, SVM provides a *cost parameter*, C, which allows the user to control the trade-off between training errors and strict margins. The increase in the value of C leads to the rise of the cost of misclassified data samples, but it causes the construction of a more precise hyperplane which may not cope with unseen samples. The selection of C and the optimization of SVM classifier is another topic, and it will be discussed in Section 3.4.1.2.

In our experiments, we used SVM^{light} implementation of SVM in C, which is developed by Thorsten Joachims [68], and available at `http://svmlight.`

joachims.org/.

3.4.1.2 Cross-validation

In the previous section, we have mentioned that the selection of cost parameter C highly affects the classification accuracy of SVM. For the accurate localization of C, we can statistically analyze our training set and estimate error in generalization (estimate how our classifier is capable of predicting data that it is not trained for) by using k-fold cross-validation. For the k-fold cross-validation, the training set is partitioned into k distinct subsets. One of these k subsets is chosen as the test set, and the remaining ones are used as training data. The accuracy of the classifier is validated with the predefined C parameter using this particular test set, and this training/validation process is repeated for each distinct subset. The average classification accuracy of these trials is used to determine the effectiveness of this particular C value. For a reasonable interval, the accuracies acquired from the use of these C values are used to find the maximum accuracy, and the corresponding C value. In our experiments, we chose to use 10-fold cross-validation, which adequately partitions our training data and does not bring high computational complexity.

3.4.2 Feature reduction

The features we have defined may provide our classifier the ability to classify the test samples at high rates. But we may encounter a problem which occurs with the increase in dimension of the feature vector and redundancy. With the growing numbers of features, the classifier starts not being able to create a successful model which maps the feature values to their corresponding class labels. This phenomenon is called *curse of dimensionality*. The curse of dimensionality brings computational complexity and it generally lowers the classification rates.

To address this problem, we may use several feature reduction algorithms. In this thesis, we employ the following methods to decrease the number of features

and select the most representative ones:

3.4.2.1 Principle component analysis

Principle component analysis (PCA) is a common and simple statistical procedure used in various fields. PCA provides the methodology to break a set of complex and correlated feature set into a lower dimension. The resulting reduced set becomes a set of uncorrelated features. The transformed features are called *principal components*.

The goal of PCA is to find the most meaningful basis for the feature set, and to redefine the features based on this basis. Simply we must find a basis P for the transformation of our original feature vector X, which will satisfy the equation $Y = PX$, where Y is the new representation of the feature vector. For this purpose, the eigenvalue decomposition of the co-variance matrix of the feature values is calculated. We have applied the cross product on our feature vector by combining the eigenvectors, starting from the eigenvector with the highest eigenvalue. The resulting data, the projection of our feature vector on the eigenvectors, are then used for classification. We used PCA in order to understand how correlated our features are. Results are given in Section 4.2.2.2.

3.4.2.2 Forward selection

Forward selection is one of the main procedures for the automated selection of variables. In forward selection, a single feature is selected and tested each time for the inclusion in the final feature subset. Starting with an empty set of features, we use each single feature in the training process of the classifier and use 10-fold cross validation set classification accuracy as a measure to or not to include the feature to the feature set. The feature which provides the highest classification accuracy is added to the feature set, and excluded from the candidate set. The remaining candidate features are tested in the next iteration, and if the feature combination with the highest accuracy exceed the achievement of the previous feature set (or

less than the previous one at most some value of ϵ), this combination is used as the base in the next iteration. This process continues until none of the combinations comply with aforementioned conditions.

The advantage of the forward selection is that it includes the best representative and discriminative features one-by-one and eliminates the ones with lower discriminative power. We make use of 10-fold cross validation for each feature subset combination in each iteration for the selection of the cost parameter C for its SVM classifier, and then we train this SVM classifier with the selected C and the corresponding features.

3.4.2.3 Backward elimination

Backward elimination mirrors the former algorithm, forward selection. In backward elimination, contrary to the forward selection, the process is started by selecting the full feature set and excluding one feature at a time. The best feature subset, which provides higher accuracy, points to a feature that is less important than the rest, because the excluded feature has decreased the accuracy less than the others (if the accuracy acquired with the corresponding subset is higher than that of the previously used feature subset). Removing low performing features in each iteration (choosing the subset with the highest accuracy and keeping it in the next step), this stepwise procedure generally converges to a highest accuracy available at some step, and then (with the removal of required features), the accuracy starts to diminish. On the other hand, removal of any one of the features may not improve the accuracy, so the algorithm does not remove any feature. Briefly, this algorithm runs until no features improve the accuracy further.

For the feature analysis with backward elimination, we have found the cost parameter C of SVM with the same process as in the case of forward selection; 10-fold cross validation is utilized to find C in each iteration for each individual feature subset.

Chapter 4

Experimental Results

This chapter is devoted to the evaluation of our experiments on histopathological colon biopsies. The preparation of our data set, inners of parameter selection, the way that the correlation between the features are analyzed, the success of our method, and comparisons are detailed in sequence.

4.1 Experimental Setup

The data set we used in our experiments consists of 213 biopsy samples taken from 58 individual patients, which are selected and collected from the cancer records of the patients during the years 2004-2007 in Hacettepe University[1]. The biopsies are stained with the hematoxylin-and-eosin technique, which is routinely used to stain biopsies in hospitals, and their images are acquired using a 20× microscope objective lens. 115 of these images are used as the *training set* and the remaining 98 of these as the *test set*. We employ 10-fold cross validation on the 115 images in the training set; the test samples are not used in this process at all. For the experiments, our data set has been examined and graded (if cancerous) by a pathologist. This process is repeated at different times to reduce

[1]The samples are collected from Department of Pathology of Hacettepe University, Ankara, Türkiye.

the intra-observer variability. The number of healthy, low-grade cancerous, and high-grade cancerous samples in these data sets are presented in Table 4.1:

	Healthy	Low-grade cancerous	High-grade cancerous	Total	Individual patients
Training set	38	37	40	115	29
Test set	34	35	29	98	29

Table 4.1: Number of samples in the training and test sets

The software is implemented as a blend of MATLAB, Java, and ANSI C code, where applicable. The software language used in each individual step is given in Appendix A. The experiments are conducted on a server with a GNU/Linux operating system, two quad-core Intel Xeon CPUs, 4 GB of DDR2 memory, three 170GB/15K SCSI hard disk drives configured as RAID 0.

4.2 Results

4.2.1 Parameter selection

In our experiments, the aim is to select the class (healthy, low-grade, high-grade) of a given biopsy sample with the best possible accuracy. For this purpose, it is fundamental to set input parameters of each step to their optimum values. The description of parameters and their effects on accuracy, how they are examined and the optimum value is found is given below:

4.2.1.1 Number of clusters k

In the k-means clustering step, the decision of appropriate k value, which is the number of clusters, yields in better segmentation of white, nuclei and stromal regions' pixels of the biopsy images. In our experiments, the decision of $k = 3$

is sufficient enough to discriminate these three regions. Usage of k with greater numbers does not increase discriminative power of the k-means clustering step. The reason is that these three regions generally have well-differentiated color distributions, even though the hematoxylin-and-eosin staining technique may not provide sharp details every time. The color distribution and conversion to Lab color space make it possible to cluster these regions with just $k = 3$. Please note that greater values of k increase the complexity and runtime of the process.

4.2.1.2 Preprocessing

Preprocessing the labeled maps may increase the classification accuracy with the removal of noise. We should analyze the impacts of the preprocessing to understand whether it is necessary or not. For this purpose, we examined how the existence of preprocessing affects the classification accuracy by a series of experiments, in which the circle-fit transform threshold value is set between 5 and 45, increasing 5 by 5. In these experiments, the multiclass classifier SVM is used with a cost parameter C of 500. The selection of this parameter value relies on the results obtained for the selection of the best C, which are shown in Figures 4.3 and 4.4. Classification results, as a function of the threshold value, are presented in Figure 4.1.

From the chart, it can be derived that preprocessing improves the classification and increases accuracy approximately ten percent for the circle-fit threshold values greater than 10. Preprocessing also preserves the acquired accuracy values at the highest peak with an accuracy of 88.78% (where the circle-fit threshold is 10). For the rest of the experiments, both preprocessed and unpreprocessed maps are used, because at the point where we achieve the highest classification rates, the discriminative power of these maps are close to each other.

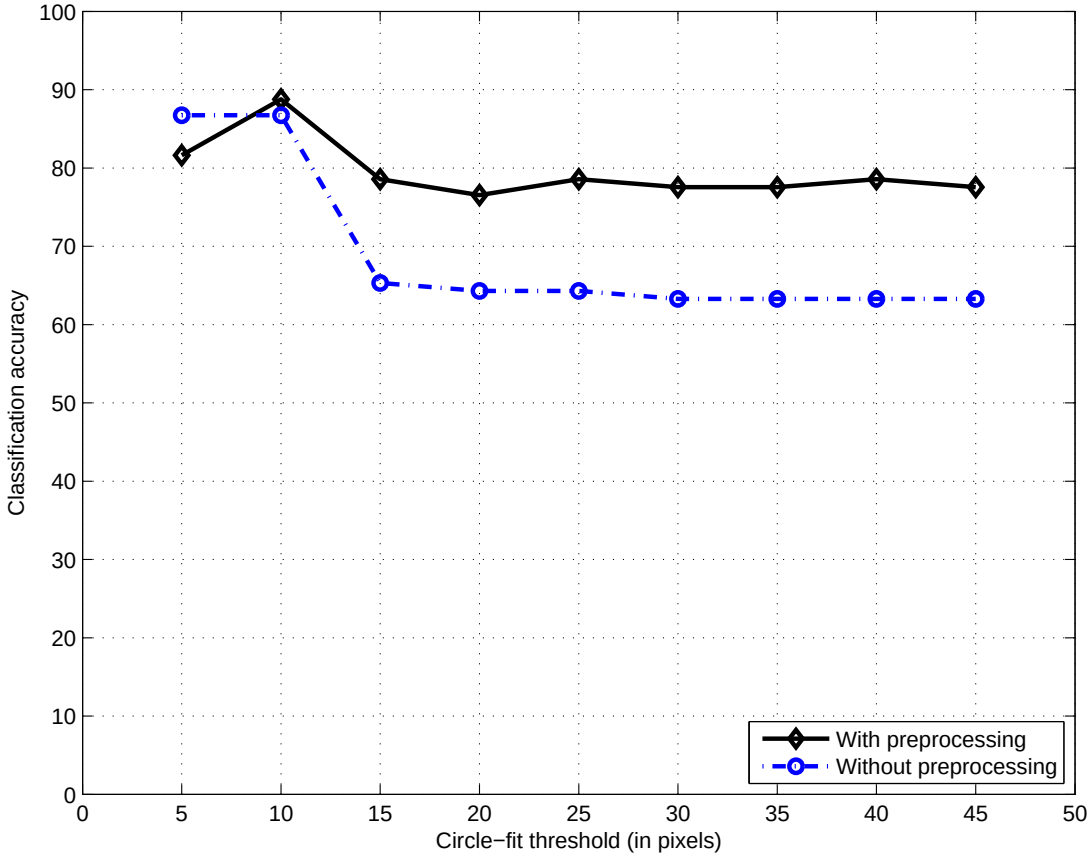
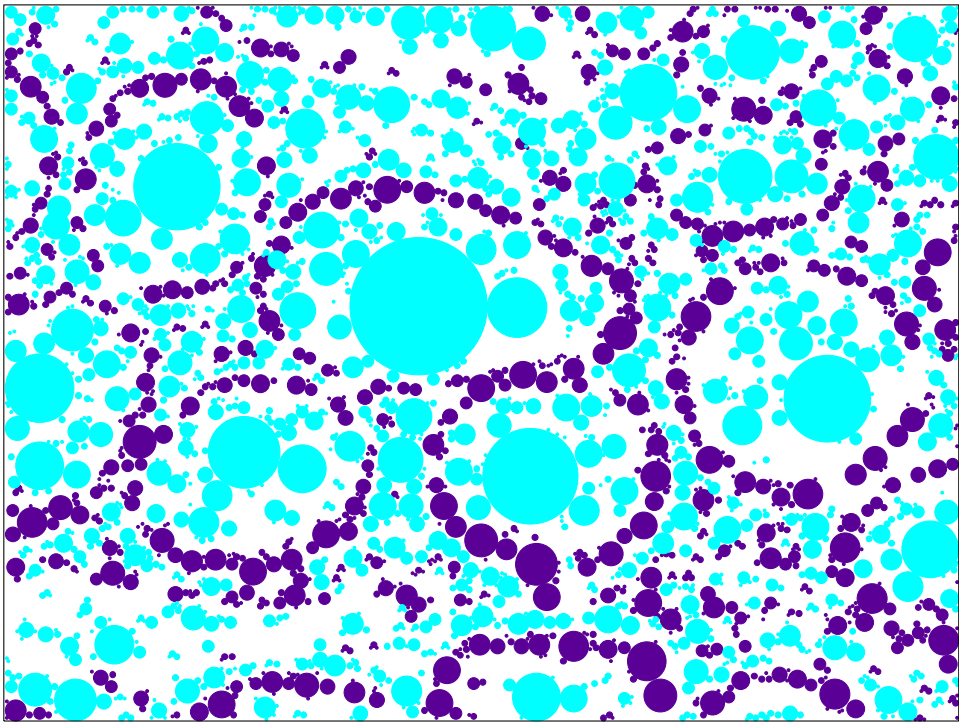


Figure 4.1: Classification results with and without preprocessing. These results are obtained by fixing the SVM cost parameter C to 500.

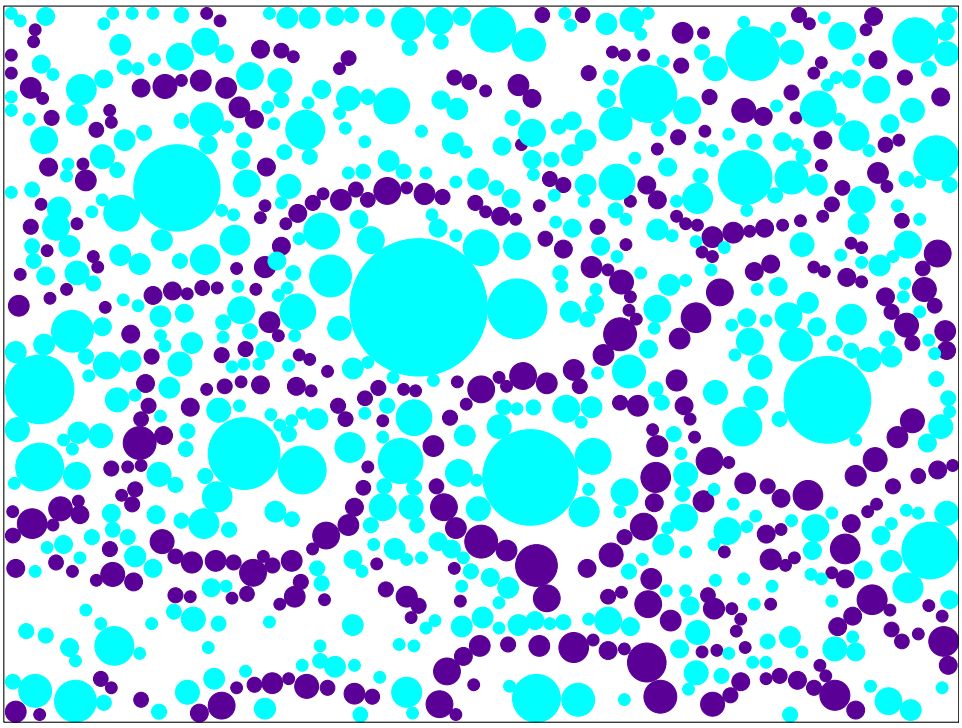
4.2.1.3 Circle-fit transform threshold

In the circle-fit transform step, there exists a threshold parameter for the algorithm which causes the elimination of the circles with fewer pixels below this predefined threshold value. This elimination lessens the number of circles, thus reducing the complexity, and runtime of this step. Circles for Figure 3.2a (see Page 30) are presented in Figures 4.2a and 4.2b, which are generated with circle-fit threshold values of 5 and 45 for both luminal and nuclei regions. In these figures, white nodes are represented by cyan circles.

Tiny circles act as noise and decrease the quality of the triangulation, so this elimination may increase the classification rates. However, with larger threshold



(a) Circle-fit threshold = 5



(b) Circle-fit threshold = 45

Figure 4.2: Circle representations of the tissue image shown in Figure 3.2a

values and the omission of required circles, it may become impossible to create a substantive triangulation and make classification. The former experiment shown in Figure 4.1 exhibits that the circle-fit algorithm maximizes the overall classification accuracy at threshold values in the close range of 10. At the highest values, the accuracy seems to increase, but at threshold value 50, most of the nuclei circles disappear and it becomes impossible to create a triangulation in some of the samples. All of these outputs are also included in the final experiments to understand the nature of the processes better.

4.2.1.4 SVM classifier and its cost parameter C

After the extraction of features, various classifiers can be chosen to examine the data: naive Bayes classifier, support vector machines, decision trees, neural networks, and so on. In our earlier experiments, we have seen that support vector machines with linear kernels work better than aforementioned classifiers for our data. However, with the selection of SVM as the classifier, another problem arises: The cost parameter C highly affects the classification accuracy. We first decided to find an optimum value for C, and to this end, we trained the classifier directly with the use of training set and varying values of C, and test the accuracy on the test set. In Figure 4.3, the effects of varying C values between 1 and 10000 against the data set obtained with different circle-fit threshold values (5, 10, and 45) are presented.

The results lead us to analyze the peak values of the classification at the beginning of the curve, where cost parameter C takes values between 1 and 2000. Figure 4.4 exhibits that the most successful classification rate of 88.78% is obtained with the circle-fit threshold value of 10 and $c \approx 500$.

The results also point out the deviation of accuracy with the change in cost parameter C and threshold values. They look promising, but present their own challenge. For each training and test set combination, a predefined C value may not strengthen the classifier, because the accuracy highly fluctuates. Thus, we can conclude that it is not rational to constrain or fix these values. Therefore, we

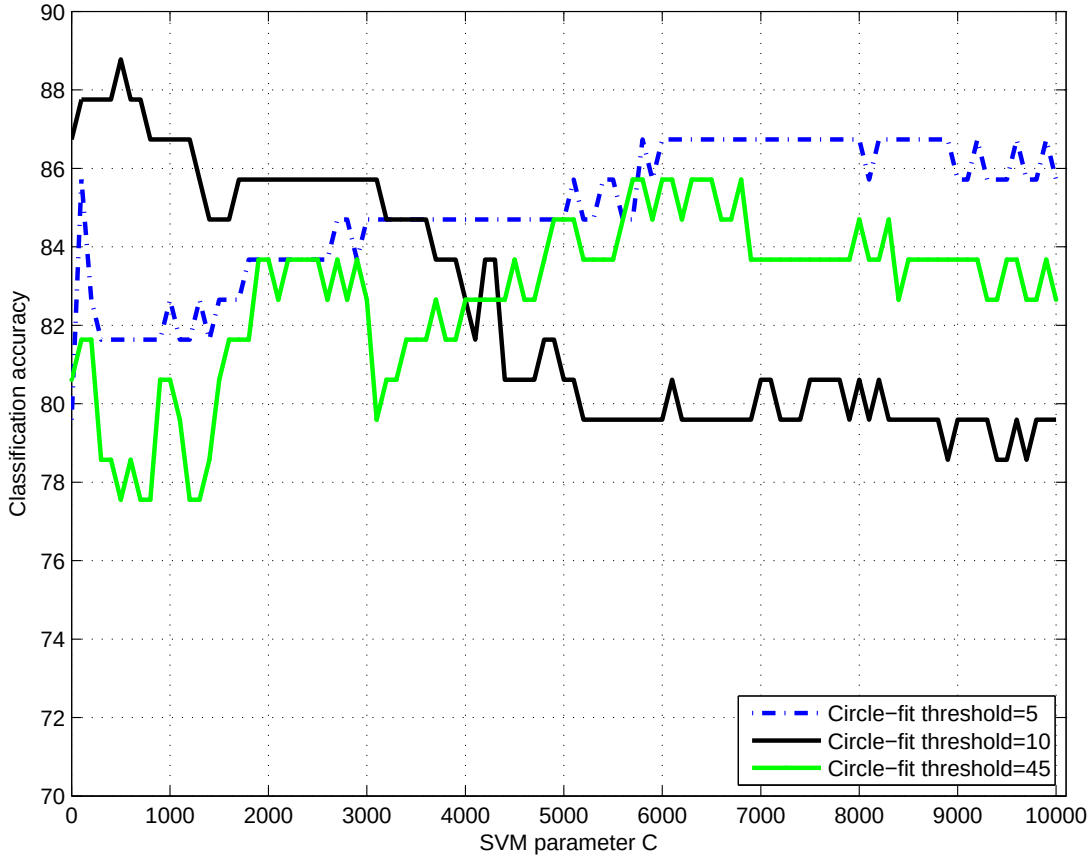


Figure 4.3: Classification results for the test set with varying values of SVM cost parameter C between 1 and 10000. These results are obtained with preprocessed data.

make use of 10-fold cross-validation to conceive the most appropriate C value for the use of SVM classifiers in the classification process. We applied 10-fold cross-validation on the training set (115 biopsy images), which is randomly divided to 5 sets of size 12 and 5 sets of size 11. In each set, there exist nearly the same amount of healthy, low-grade cancerous, and high-grade cancerous samples. In each iteration, nine of these sets are used as training data, and the remaining single set is retained as the validation set. The accuracy is calculated by averaging the accuracy results of all $K = 10$ trials. The C value which provides the most reliable classification, the one with the maximum classification accuracy, is used in the training of the classifier over the training set and the classification results of the test set is acquired.

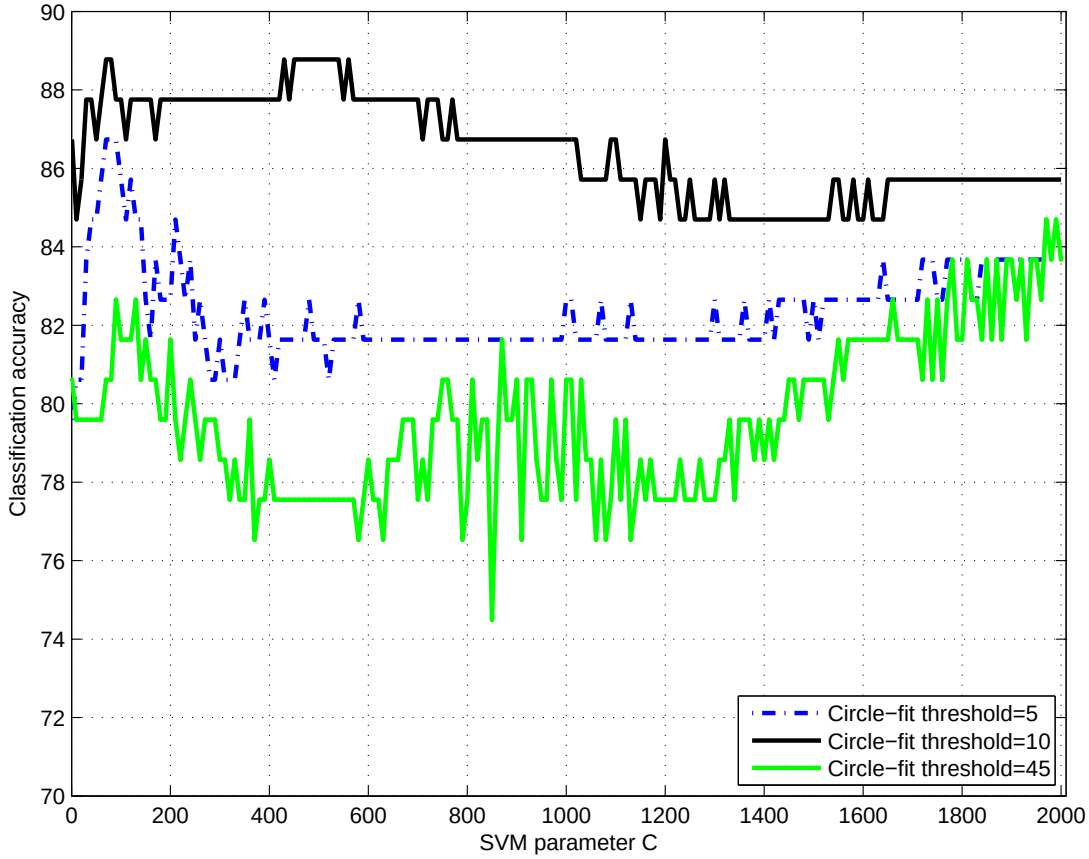


Figure 4.4: Classification results for the test set with varying values of SVM cost parameter C between 1 and 2000. These results are obtained with preprocessed data.

4.2.2 Feature selection and reduction

In the previous section, we discussed how the parameters are analyzed and fixed. Even with the best parameter selection, classification accuracy could be insufficient. The reason is that some features may be malicious, and they may not be concerned with the health status of the tissues. Besides, some of these features may also correlated with some other features, and with the increase in feature space dimension, the classifier may not work well as expected with the introduction of curse of dimensionality. In this section, we will be investigating the techniques on feature selection and reduction. We will analyze the information obtained from forward selection and backward elimination results, and also experiment on a set of preselected features.

4.2.2.1 Manual selection of features

We have decided to use a subset of the predefined features to improve the quality of classification process and eliminated some features. For the reason, we analyzed the features and chose the subset below:

1. *Average degree, number of isolated nodes, number of end nodes*: Degree of a node is a local feature and, in our case, it reflects the connectivity properties of each individual cell. On a larger scale, the average degree, number of isolated nodes and number of end nodes features help us understand the connectivity of the tissue. For example, with the development of cancer, regular structure of luminal components suffer a severe loss and cells get separated from each other. This fact affects the average degree and number of isolated nodes in the constrained Delaunay triangulation directly.

2. *Average clustering coefficient for nodes with* $d(p_i) \geq 2$ *, diameter, average edge length*: These features go beyond the local connectivity and present us the information about the neighborhood characteristics of the nodes in our graph, the constrained Delaunay triangulation. We did not use average clustering coefficient directly, because the nodes with no or one neighbor are already symbolized by number of isolated nodes and number of end nodes.

3. *Average triangle area, standard deviation of triangle area*: Cell groups around the luminal regions get thickened with the development of a low-grade cancer. However, the structure is completely distorted in high-grade cancerous tissues, and the local connectivity of the cells is lost. Only some cell groups may form a line of cells and some thin, long triangles are formed in these regions. The structuring of the tissues can be represented by the triangles of the constrained Delaunay triangulation and the variation in these tissues can be measured with these properties. We have added the features related to the triangulation to our predefined feature set for this reason.

With the use of these features and constrained Delaunay triangulation, we have acquired the classification accuracies that are shown in Tables 4.2 and 4.3 for the training and test set, respectively. These experiments are carried out with circle-fit threshold value of 10 and preprocessed images. Here, the SVM parameter C is selected using 10-fold cross validation. Tables 4.4 and 4.5 demonstrate the accuracies acquired with the use of features extracted from Delaunay triangulation. These results show that constrained Delaunay triangulation improves the classification at a rate of 2.61 for the training set and 12.24 for the test set.

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	30	8	0	78.95
Low-grade	0	27	10	72.97
High-grade	0	5	35	87.50
Overall Accuracy				80.00

Table 4.2: The confusion matrix and the training set classification accuracy of the constrained Delaunay triangulation approach with the circle-fit threshold value being selected as 10. These results are obtained with using preprocessed images.

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	30	1	3	88.24
Low-grade	0	28	7	80.00
High-grade	0	2	27	93.10
Overall Accuracy				86.73

Table 4.3: The confusion matrix and the test set classification accuracy of the constrained Delaunay triangulation approach with the circle-fit threshold value being selected as 10. These results are obtained with using preprocessed images.

4.2.2.2 PCA

We have used PCA to find out the prevalence of correlation within our feature set. For this purpose, we have sorted the eigenvectors V of the correlation matrix of our feature set according to the magnitude of their eigenvalues λ_i

$$V_1 > V_2 > \cdots > V_D \, (with \; \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D)$$

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	33	2	3	86.84
Low-grade	3	19	15	51.35
High-grade	0	3	37	92.50
Overall Accuracy				77.39

Table 4.4: The confusion matrix and the training set classification accuracy of the Delaunay triangulation approach with the circle-fit threshold value being selected as 10. These results are obtained with using preprocessed images.

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	29	0	5	85.29
Low-grade	3	20	12	57.14
High-grade	1	4	24	82.76
Overall Accuracy				74.49

Table 4.5: The confusion matrix and the test set classification accuracy of the Delaunay triangulation approach with the circle-fit threshold value being selected as 10. These results are obtained with using preprocessed images.

Subsequently, we have multiplied our feature set with the first k eigenvectors $V_1, V_2, \cdots, V_k$ and reduced the dimension from D to k. The results obtained with the different values of k are given in Figures 4.5 and 4.6 for the training and test set, respectively.

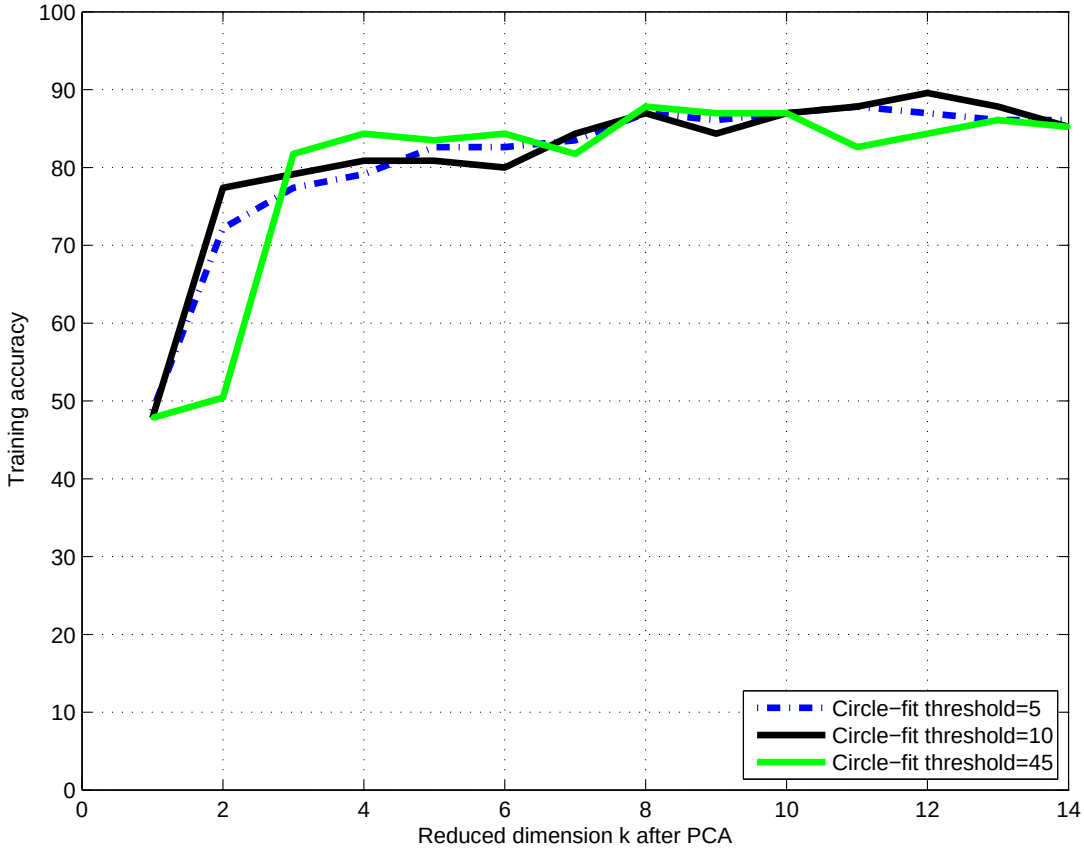


Figure 4.5: Classification results in PCA for the training set. These results are obtained by choosing the SVM cost parameter C individually with 10-fold cross-validation in each iteration and with preprocessed images.

The results expose the fact that the use of 10-fold cross-validation optimizes the results for the training set and the accuracy increases regularly with the increase in the value of k. The training accuracy reaches its peak value (89.57%) when $k = 12$ with circle-fit threshold value of 10. However, test set accuracy does not follow the behaviors of the training set accuracy, and produces an unstable curve. The test set accuracy is only 79.59% where the training set earns its maximum.

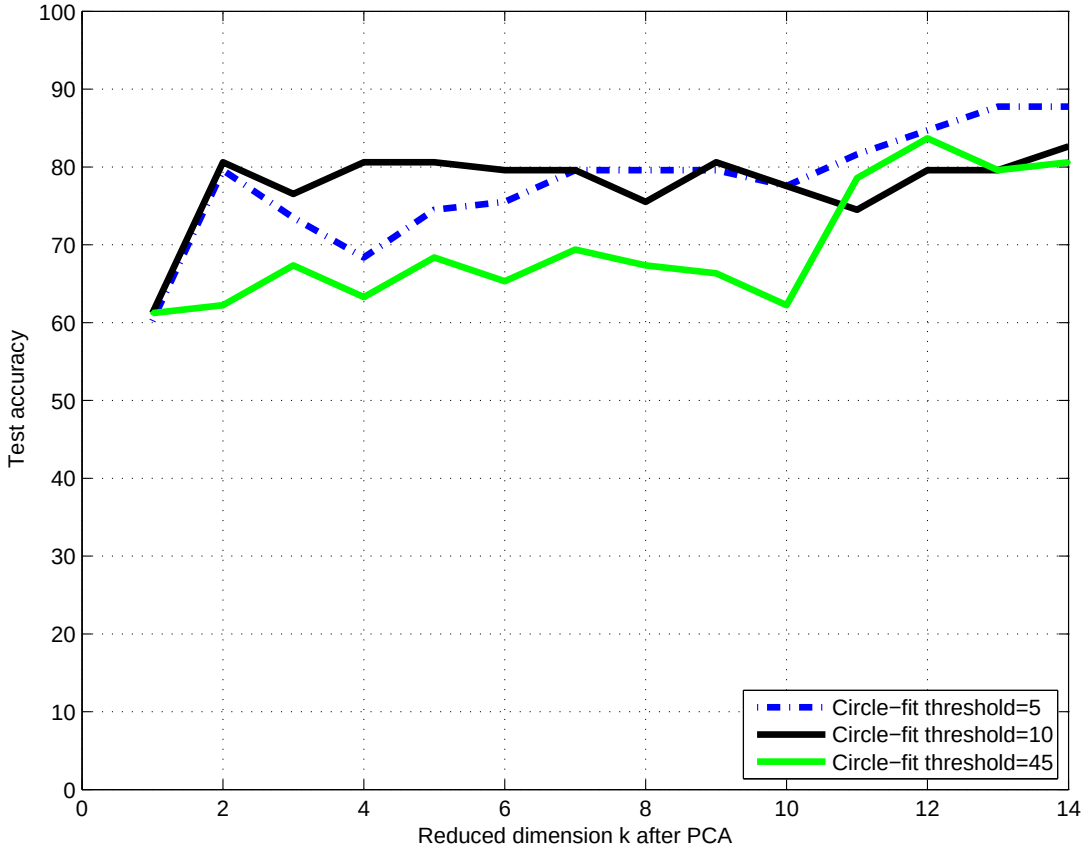


Figure 4.6: Classification results in PCA for the test set. These results are obtained by choosing the SVM cost parameter C individually with 10-fold cross-validation in each iteration and with preprocessed images.

4.2.2.3 Forward selection

Both manual selection and PCA gives us an idea that we can use the forward selection. Likewise, for the detection of C parameter of SVM, 10-fold cross-validation is used. After selecting the value of C, the classification accuracy for both training and test sets are calculated. The decision on whether to continue the iteration or to stop is based on the accuracy of 10-fold cross-validation set. Table 4.6 demonstrates the classification results for the 10-fold cross-validation. In Tables 4.7 and 4.8, classification accuracies with the selected C values are demonstrated for the training set and the test set, respectively.

10-fold cross val.		Iteration								
		#1	#2	#3	#4	#5	#6	#7	#8	#9
Circle-fit threshold	5	56.94	63.86	77.73	78.56	78.64	79.47	81.20	81.97	–
	10	58.42	76.67	79.12	82.80	86.21	87.12	87.12	–	–
	15	56.82	61.82	65.30	66.20	66.36	67.12	68.11	68.26	68.26
	20	56.82	61.60	–	–	–	–	–	–	–
	25	56.82	61.60	–	–	–	–	–	–	–
	30	56.82	61.60	–	–	–	–	–	–	–
	35	56.82	62.80	65.56	–	–	–	–	–	–
	40	57.70	64.55	65.30	65.45	65.53	66.52	69.92	71.73	–
	45	56.74	64.14	64.31	–	–	–	–	–	–

Table 4.6: Forward selection results for 10-fold cross-validation

For the circle-fit threshold value of 10, which helps us acquire the highest classification accuracy, the accuracy results for 10-fold cross-validation, training and test sets are also given in Figure 4.7.

From these and the previous results, we can derive that the best performing circle-fit threshold value is 10. At this value, the selected features and the test set accuracies at the corresponding iterations are given in Table 4.9. In the first iteration, the selected feature “end node number” itself may provide a classification accuracy of 55.10%, and with the use of other features, the accuracy can reach

Training set		Iteration								
		#1	#2	#3	#4	#5	#6	#7	#8	#9
Circle-fit threshold	5	52.17	60.87	77.39	76.52	79.13	81.73	82.61	82.61	–
	10	57.39	75.65	79.13	83.48	86.96	86.96	86.96	–	–
	15	48.70	62.61	64.34	67.83	69.56	66.95	68.70	72.17	73.91
	20	48.70	63.48	–	–	–	–	–	–	–
	25	48.70	63.48	–	–	–	–	–	–	–
	30	48.70	63.48	–	–	–	–	–	–	–
	35	48.70	60.87	63.48	–	–	–	–	–	–
	40	48.70	60.87	62.61	66.95	62.61	71.30	71.30	74.78	–
	45	52.17	59.13	66.95	–	–	–	–	–	–

Table 4.7: Forward selection results for the training set

Test set		Iteration								
		#1	#2	#3	#4	#5	#6	#7	#8	#9
Circle-fit threshold	5	46.94	64.29	63.37	73.47	72.45	69.39	70.41	68.37	–
	10	55.11	78.56	82.65	79.59	80.61	80.61	80.61	–	–
	15	36.72	52.40	63.27	65.31	66.33	63.27	66.33	66.33	67.34
	20	36.72	53.60	–	–	–	–	–	–	–
	25	36.72	53.60	–	–	–	–	–	–	–
	30	36.72	53.60	–	–	–	–	–	–	–
	35	36.72	52.40	66.33	–	–	–	–	–	–
	40	36.72	52.40	64.29	62.24	62.24	68.37	68.37	69.39	–
	45	50.00	55.10	60.20	–	–	–	–	–	–

Table 4.8: Forward selection results for the test set

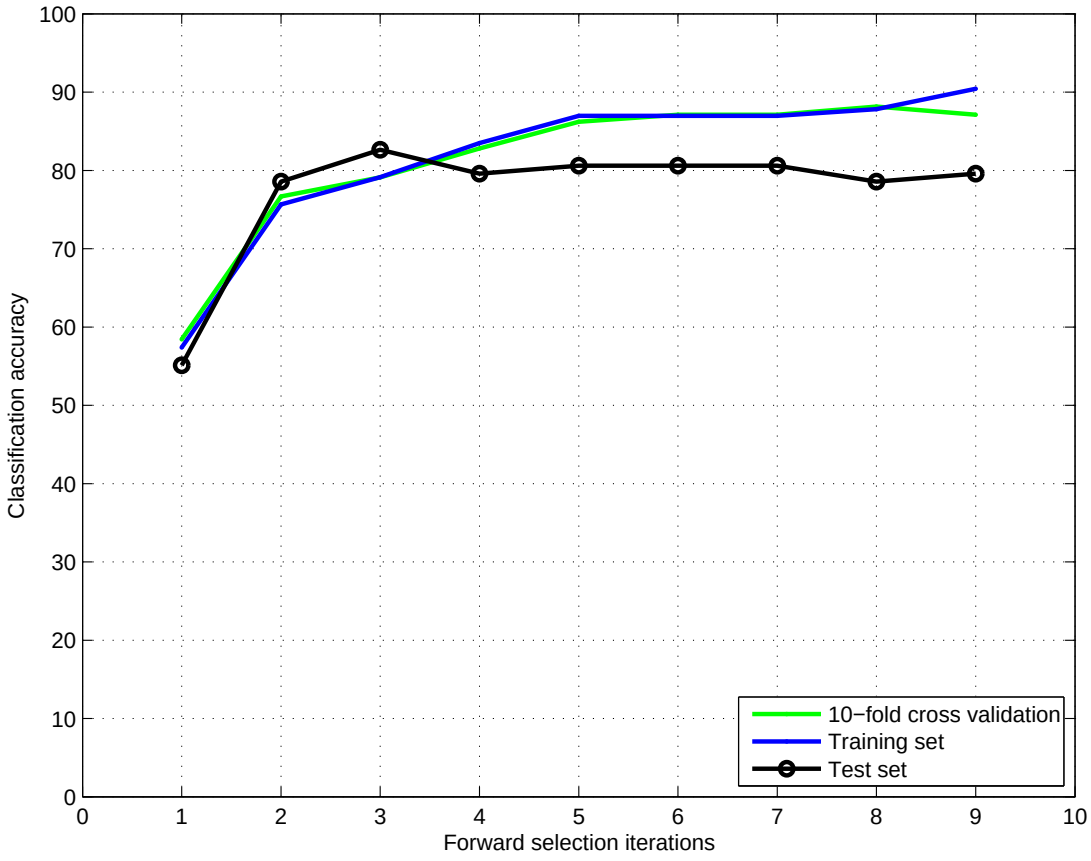


Figure 4.7: Classification results in forward selection. These results are obtained by choosing the SVM cost parameter C individually with 10-fold cross-validation in each iteration and with preprocessed images.

approximately upto 80%. However, when we examine the results presented in Figure 4.1, we can see that with the use of all features, preprocessing, and circle-fit threshold value of 10, we can easily reach 86.73%. The drawback of forward selection is that it does not consider the collective power of feature subsets [11]. The addition of a feature may turn an existing selected feature into a non-significant one. For example, any feature a and b may provide less idea about the structure by themselves, considering the contribution of remaining features, but the subset composed of these two features a and b may bring out the highest classification accuracy in the subset collection of feature couples. Therefore, we attempted to test another feature reduction method, backward elimination.

Iteration	**Selected feature**	**Classification accuracy**
#1	*End node number*	55.10%
#2	*Standard deviation of triangle area*	78.57%
#3	Average degree for nodes with $d(p_i) \geq 2$	82.65%
#4	Standard deviation of edge length	79.59%
#5	Average eccentricity	80.61%
#6	Number of components with $n \geq 2$	80.61%
#7	*Average degree*	80.61%
#8	*Isolated node number*	78.57%
#9	*Average triangle area*	79.59%

Table 4.9: Selected features and the corresponding test set accuracies in forward selection. Our manually selected features are written in italics.

4.2.2.4 Backward elimination

The resulting classification accuracy values are given in Tables 4.10, 4.11, and 4.12, for 10-fold cross-validation, training set, and test set, respectively.

The disadvantage of backward elimination is its dependence of large number of feature subsets. But, in our case, the number of features are small that this drawback does not become the main disadvantage. It also has the same disadvantage of forward selection, that is it does not consider the unified effects of eliminated features, so one or more of the dropped features may become significant if added

10-fold cross val.		Iteration				
		#1	#2	#3	#4	#5
Circle-fit threshold	5	81.6	–	–	–	–
	10	85.46	86.29	87.20	88.11	–
	15	69.17	69.17	–	–	–
	20	69.23	–	–	–	–
	25	69.23	–	–	–	–
	30	69.92	–	–	–	–
	35	69.92	–	–	–	–
	40	69.92	70.76	70.83	71.67	72.50
	45	70.00	70.76	71.67	72.50	72.50

Table 4.10: Backward elimination results for 10-fold cross-validation

Training set		Iteration				
		#1	#2	#3	#4	#5
Circle-fit threshold	5	86.9	–	–	–	–
	10	88.70	89.57	90.44	92.17	–
	15	77.39	77.39	–	–	–
	20	73.91	–	–	–	–
	25	73.91	–	–	–	–
	30	72.17	–	–	–	–
	35	72.17	–	–	–	–
	40	73.40	73.4	72.17	73.91	74.78
	45	73.40	73.40	73.40	73.91	74.78

Table 4.11: Backward elimination results for the training set

Test set		Iteration				
		#1	#2	#3	#4	#5
Circle-fit threshold	5	79.59	–	–	–	–
	10	83.67	87.76	81.63	80.61	–
	15	66.33	66.33	–	–	–
	20	67.34	–	–	–	–
	25	67.34	–	–	–	–
	30	67.34	–	–	–	–
	35	66.33	–	–	–	–
	40	65.31	65.31	64.29	65.31	63.27
	45	64.29	65.31	66.33	64.29	67.34

Table 4.12: Backward elimination results for the test set

Iteration	**Selected feature**	**Classification accuracy**
#1	Average degree	83.67%
#2	End node number	87.76%
#3	Avg. clust. coeff. for nodes with $d(p_i) \geq 2$	81.63%
#4	Giant component ratio	80.61%
#5	Diameter	79.59%
#6	Isolated node number	79.59%

Table 4.13: Eliminated features and the corresponding test set accuracies in backward elimination.

to the optimized feature subset. As in our experiments, the use of backward elimination may provide better accuracy for the experiments on cross-validation and training sets, but the highest accuracy for the test set may remain the same. Figure 4.8 shows the classification results for 10-fold cross-validation, training and test sets for the circle-fit threshold value fixed at 10. The selected features are also presented in Table 4.13 at this circle-fit threshold level.

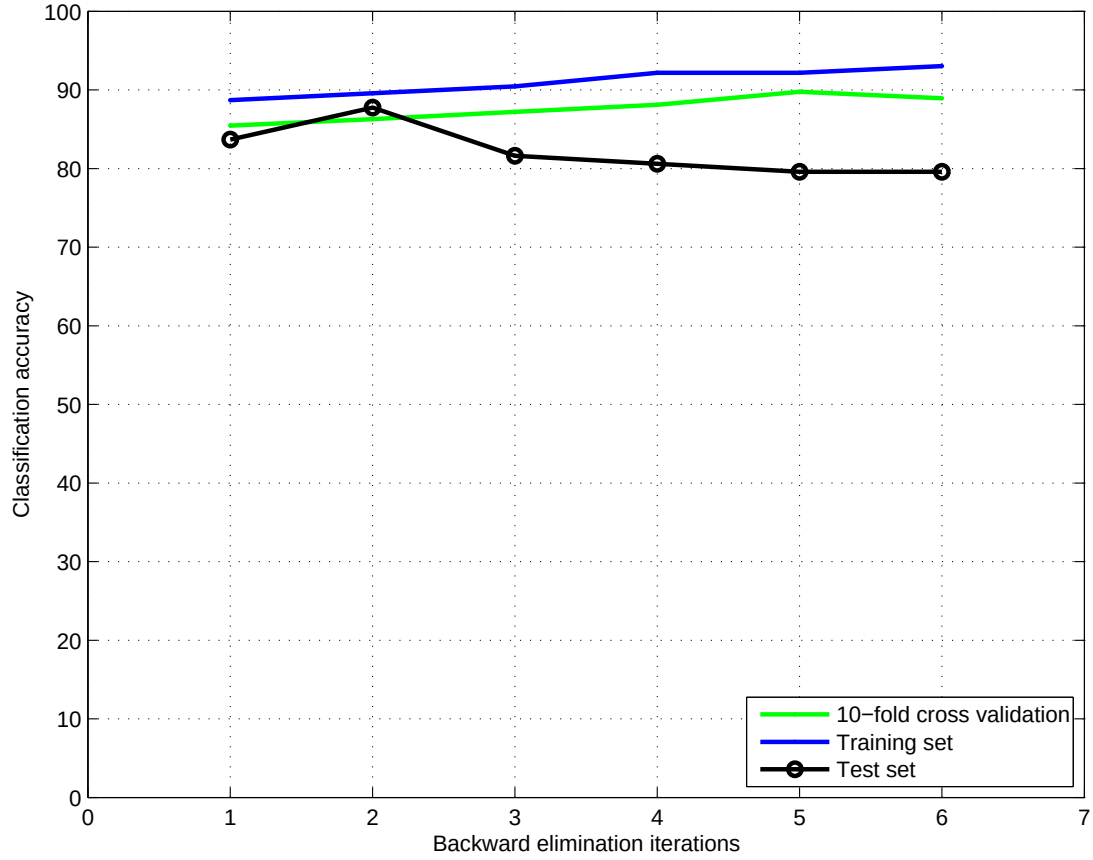


Figure 4.8: Classification results in backward elimination. These results are obtained by choosing the SVM cost parameter C individually with 10-fold cross-validation in each iteration and with preprocessed images.

4.2.3 Comparison with Delaunay Triangulation

Earlier in this chapter, we analyzed the effects of parameters and feature selection on our constrained Delaunay triangulation (CDT) and we attempted to maximize

the classification accuracy by using the features acquired from the CDT. With the maximization of CDT in terms of classification accuracy, we should examine its achievement over the classical Delaunay triangulation. Towards this end, we first used the same set of features (in fact, using a subset of features which includes the applicable ones on DT). Tables 4.14 and 4.15 present the classification accuracy of constrained Delaunay triangulation acquired with the use of samples which are preprocessed by morphological operators and processed with a circle-fit threshold of 10. Here we use all of the features defined for CDT.

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	38	0	0	100.00
Low-grade	3	26	8	70.27
High-grade	0	3	37	92.50
Overall Accuracy				87.83

Table 4.14: Training set accuracy obtained by the constrained Delaunay triangulation. The circle-fit threshold value is selected as 10.

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	31	0	3	91.18
Low-grade	2	25	8	71.43
High-grade	0	1	28	96.55
Overall Accuracy				85.71

Table 4.15: Test set accuracy obtained by the constrained Delaunay triangulation. The circle-fit threshold value is selected as 10.

From Tables 4.14 and 4.15, we can see that we reach 87.83% accuracy for the training set and 85.71% accuracy for the test set, with the use of constrained Delaunay triangulation. However, if we made use of Delaunay triangulation, the results would come out as in Tables 4.16 and 4.17. We would be reaching 77.39% accuracy for the training set and 76.53% accuracy for the test set.

Figure 4.9 compares the test set accuracies for constrained Delaunay triangulation (CDT) and Delaunay triangulation (DT), while Figure 4.10 presents the

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	33	2	3	86.84
Low-grade	3	21	13	56.76
High-grade	2	3	35	87.50
Overall Accuracy				77.39

Table 4.16: Training set accuracy obtained by the Delaunay triangulation. The circle-fit threshold value is also selected as 10.

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	29	0	5	85.29
Low-grade	4	21	10	60.00
High-grade	1	3	25	86.21
Overall Accuracy				76.53

Table 4.17: Test set accuracy obtained by the Delaunay triangulation. The circle-fit threshold value is also selected as 10.

accuracy difference between these two. From the results, we can derive that use of CDT introduces an improvement of 2.04% at minimum and 18.36% at maximum at different threshold values over the use of DT. When the circle-fit threshold is selected as 10, where both triangulations provide the highest classification accuracy, CDT provides 9.18% better rate.

We must study the effects of CDT and DT on individual classes (healthy, low-grade cancerous, high-grade cancerous) as well in order to better understand and evaluate the effects of different triangulation schemes. Figures 4.11, 4.12, and 4.13 present the test set accuracies for different classes of healthy, low-grade cancerous, and high-grade cancerous.

From Figure 4.11, we can understand that the healthy tissues are classified well by both using features extracted from CDT and DT. With the use of circle-fit threshold value of 10 and higher, the tissues are classified well enough, and at least 85% accuracy is obtained. This fact may show that either the healthy tissues are classified well or the classifier is prone to mark all test images as

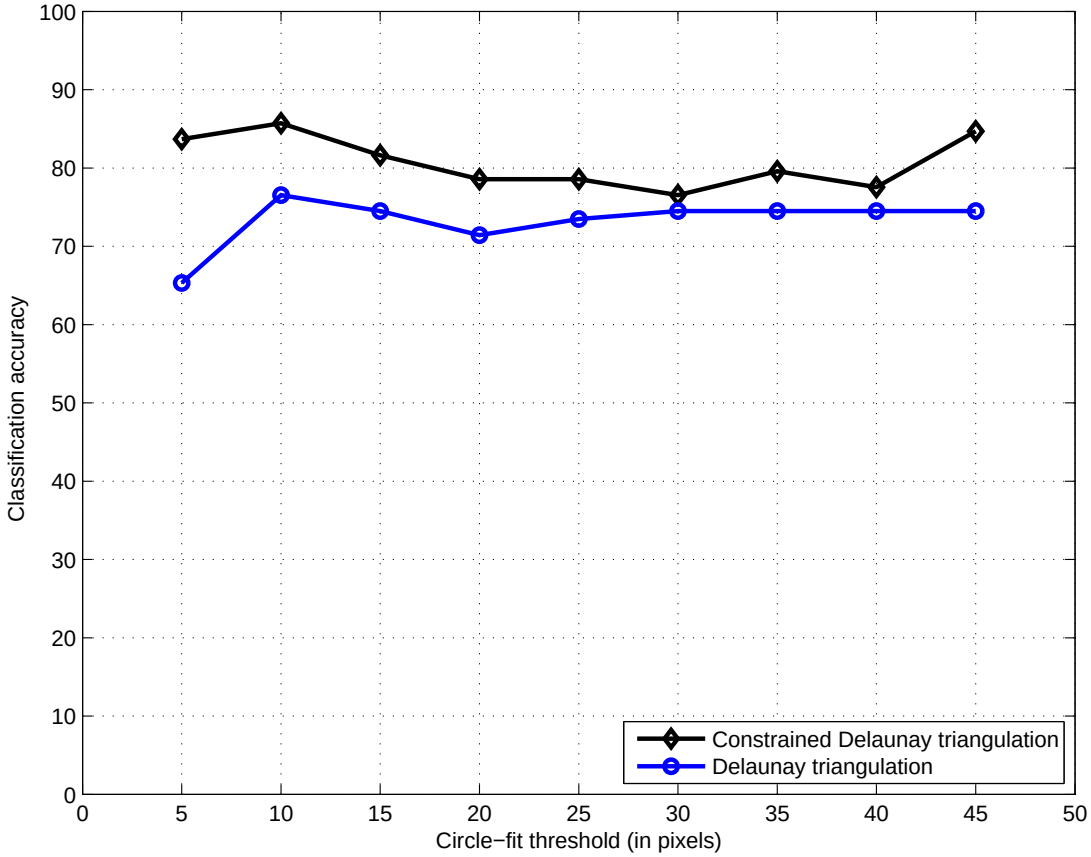


Figure 4.9: Test accuracies of constrained Delaunay triangulation and Delaunay triangulation. These results are obtained by choosing the SVM cost parameter C individually with 10-fold cross-validation in each iteration and with preprocessed images.

healthy, bringing higher classification accuracy for healthy tissues, and resulting in the loss of classification for other classes. To understand the system further, we have to look at the classification results for low-grade and high-grade cancerous tissues.

Figure 4.12 demonstrates that the low-grade cancerous tissues constitute a problem for our classification. With the use of Delaunay triangulation, at most 62.86% accuracy is obtained when the circle-fit threshold value is selected to be 5. Furthermore, when the circle-fit threshold value is set to be 10, where we reach our maximum classification performance in general, 60% of the test samples are classified correctly. However, this percentage points out that more than one-third

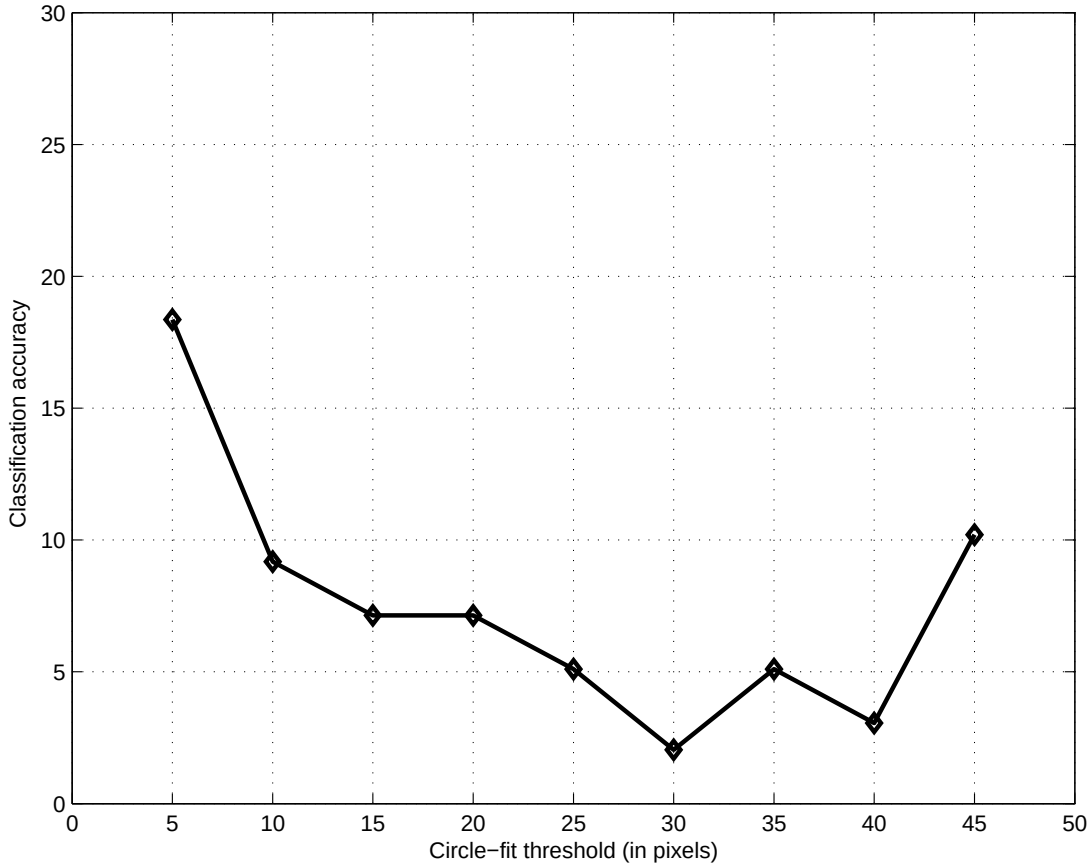


Figure 4.10: The difference in test accuracies of constrained Delaunay triangulation and Delaunay triangulation. (See Figure 4.9)

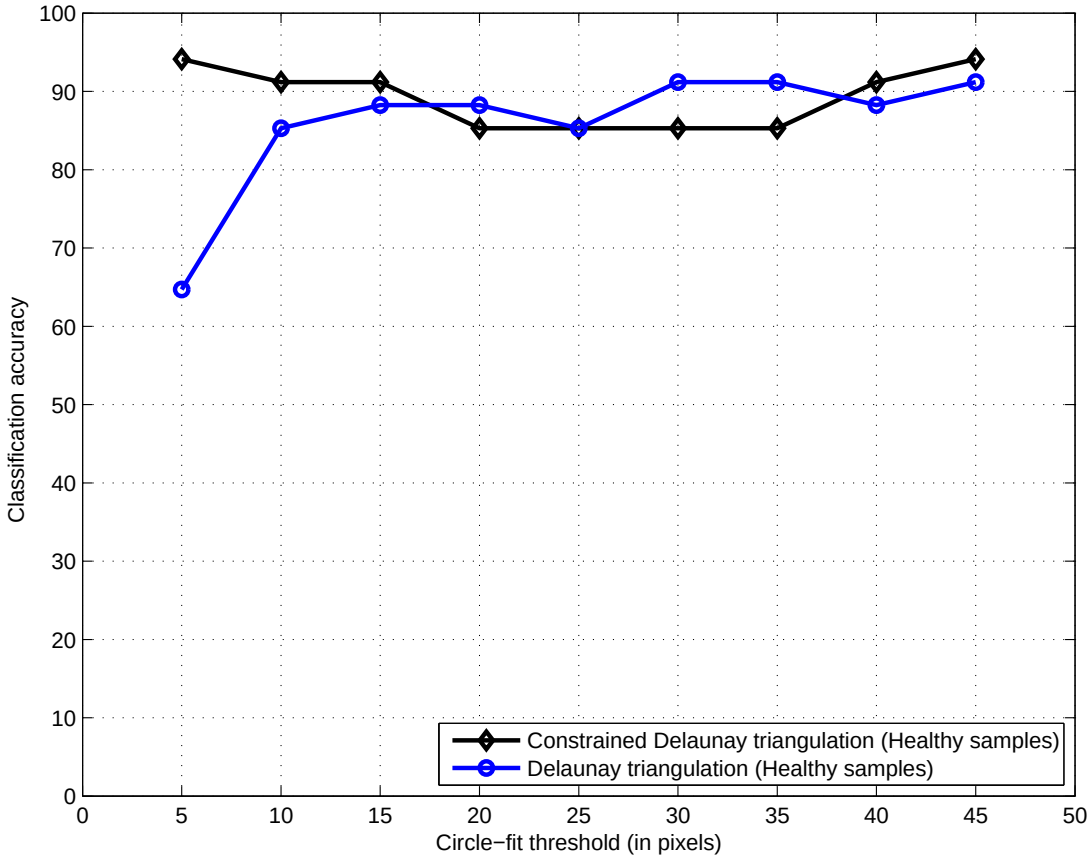


Figure 4.11: The test set accuracies obtained by constrained Delaunay triangulation and Delaunay triangulation in healthy tissues. (See Figure 4.9)

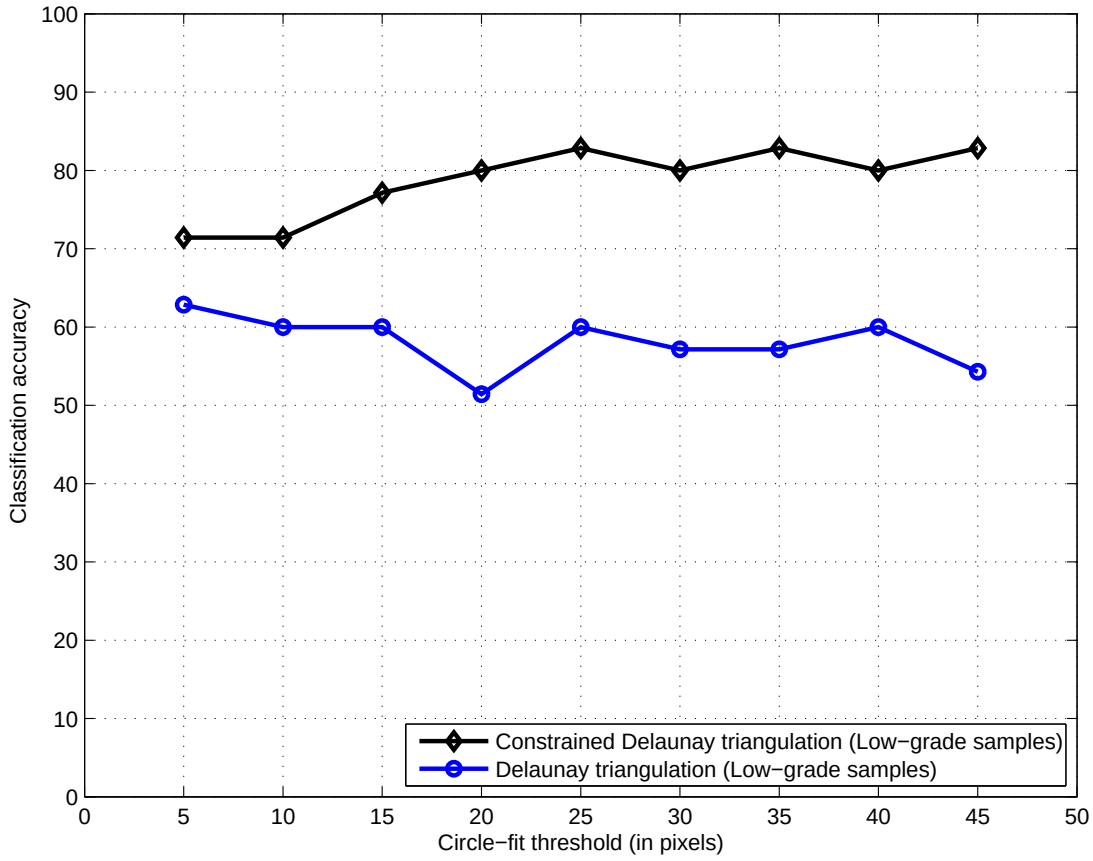


Figure 4.12: The test set accuracies obtained by constrained Delaunay triangulation and Delaunay triangulation on low-grade cancerous tissues. (See Figure 4.9)

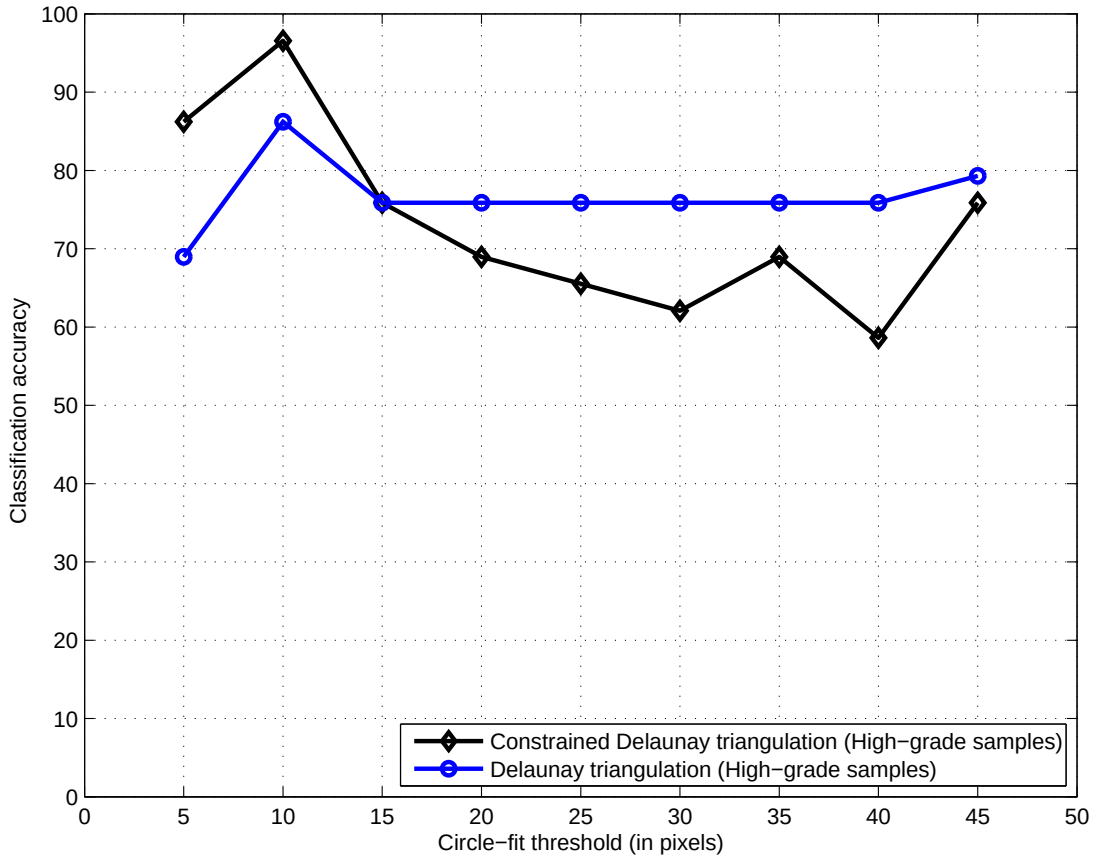


Figure 4.13: The test set accuracies obtained by constrained Delaunay triangulation and Delaunay triangulation on high-grade cancerous tissues. (See Figure 4.9)

of all low-grade cancerous samples are misclassified.

Figure 4.12 also shows that the introduction of constrained Delaunay triangulation increases the test set accuracy by 8.57%-28.57%, where circle-fit threshold values are 5 and 20/45, respectively. These results show that constrained Delaunay triangulation performs more accurately than Delaunay triangulation (20.64% increase in classification rates on average) for the classification of low-grade cancerous samples.

Last of all, we have the results acquired from the classification of high-grade cancerous samples (Figure 4.13). The results for the CDT are more unsteady than the healthy and low-grade cancerous classes. The accuracy starts from 86.21% where the circle-fit threshold is 5, jumps to 96.55%, and then decreases to 58.52%. The decrease is attributed to the accuracy increases for the other classes. The results of other classes may not be improved without this compensation. The rise in the classification rate of healthy and low-grade cancerous samples are cancelled by this great fall in the classification accuracy of high-grade cancerous samples at greater threshold values. The peak point of the rates is at threshold 10.

From these individual class observations, we can conclude that greater values of the circle-fit threshold provide us the chance to classify healthy and low-grade cancerous samples better, but result in a total loss in high-grade cancer classification. As presented in Figure 4.9, the optimum point, where we can differentiate a set of healthy, low-grade cancerous, and high-grade cancerous tissues, is the one where the circle-fit threshold value is set to be 10.

The previous results are obtained using preprocessed images. We also examine the classification success of images without preprocessing. Figure 4.14 presents the test set accuracies for constrained Delaunay triangulation and Delaunay triangulation, while Figure 4.15 shows the difference between these accuracies.

From Figures 4.14 and 4.15, we may see that the introduction of CDT improves the DT better than it did in the case of preprocessed images. It also has better classification than those in preprocessed images where the circle-fit threshold is 10. However, if we examine the rest of the results, the CDT on non-preprocessed

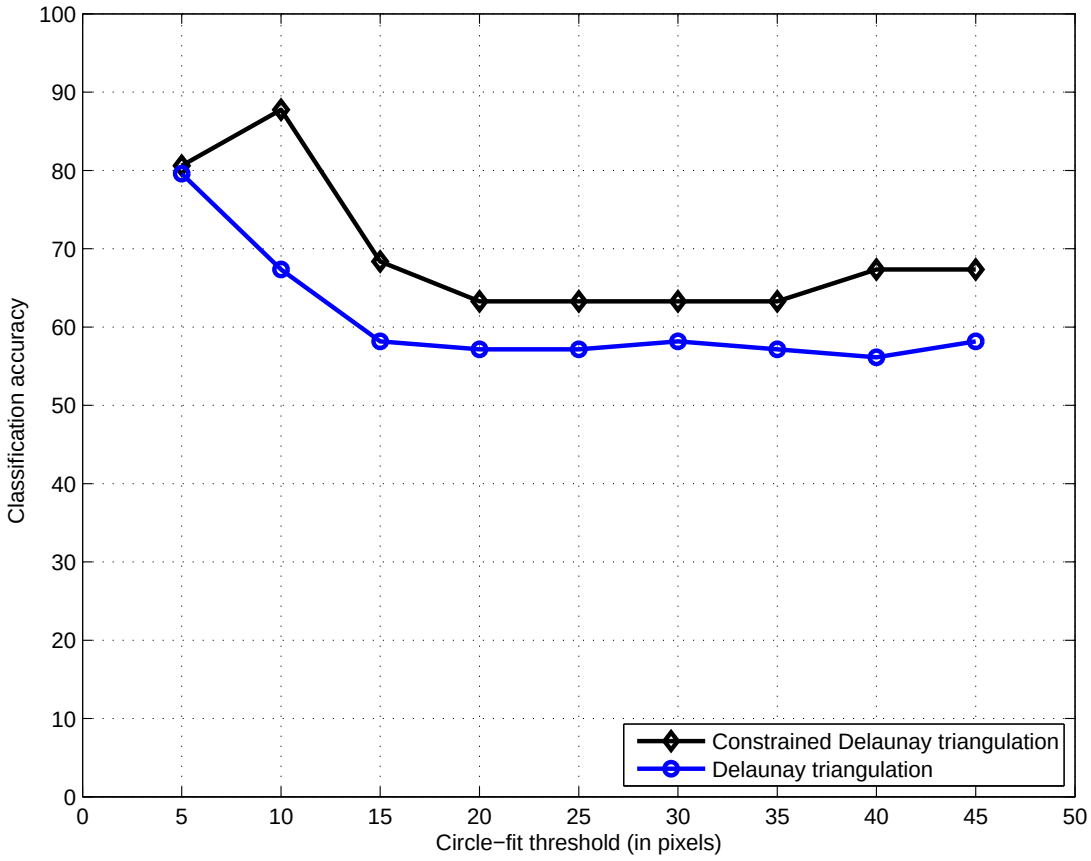


Figure 4.14: Test accuracies of constrained Delaunay triangulation and Delaunay triangulation. These results are obtained by choosing the SVM cost parameter C individually with 10-fold cross-validation and using non-preprocessed images.

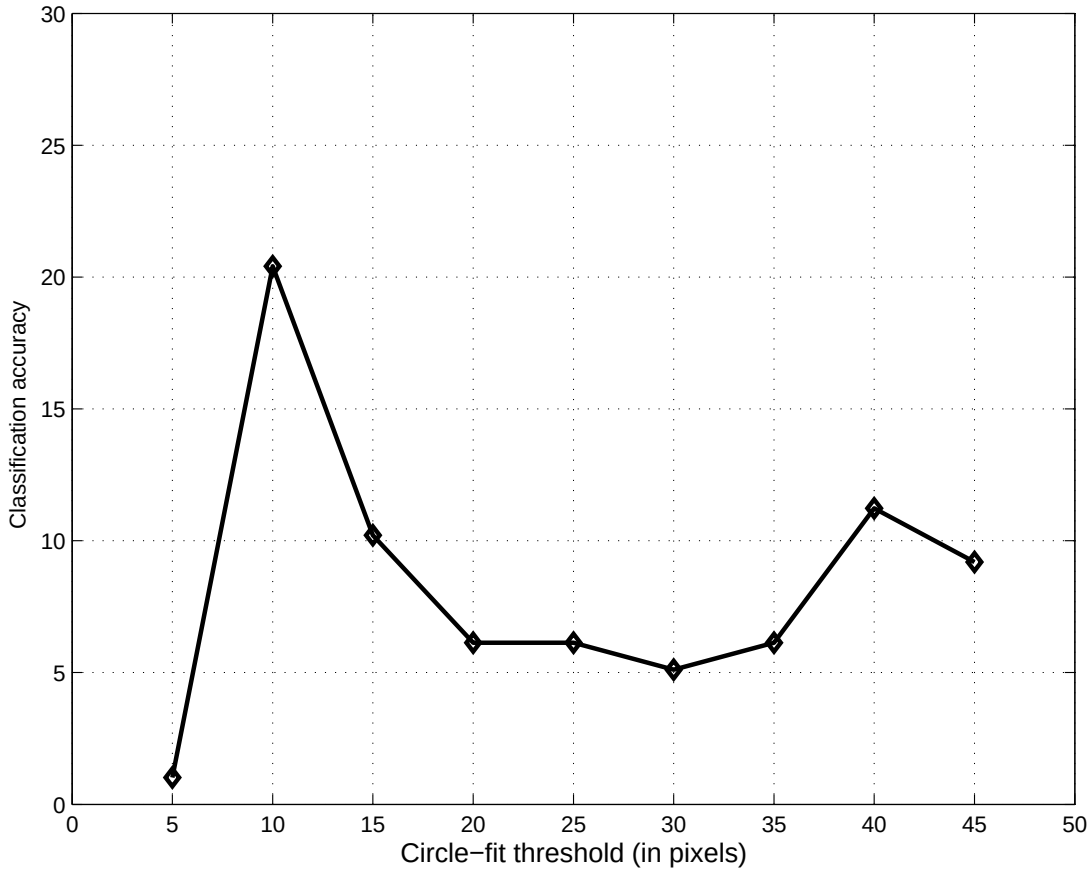


Figure 4.15: The difference between in test accuracies of constrained Delaunay triangulation and Delaunay triangulation. (See Figure 4.14)

images has obviously worse classification accuracies than those of the preprocessed ones. We conclude that the set of preprocessed images forms a more solid base for future development of the experiments.

4.3 Discussion

In this section, we will provide a discussion on the advantages and disadvantages of the algorithms we used, our feature choices, and the success of the algorithm we have developed.

4.3.1 Parameter selection

The selection of parameters is the first factor that is deeply interconnected with the increase in classification accuracy. For the k-means step, it was easy to choose $k = 3$ as the number of clusters and this selection is sufficient for the success of segmentation. However, it was difficult to choose whether to apply preprocessing or not on this clustered image maps, since they provide approximately equal accuracy at their highest peak. Our experiments helped us understand that preprocessing is better (least or more) and it provides a more solid base for the future development of our algorithm.

The circle-fit threshold is another parameter that is easy to decide on. In most of our experiments, a circle-fit threshold value of 10 turned out to be the most accurate selection. It was the most stable and reliable choice. On the contrary, the selection of cost parameter of our SVM classifier remained the biggest problem. At first, we tried to maximize the accuracy by finding a fixed C value, using the accuracy acquired from the training and test sets. However, the results had a very unstable amplitude of accuracy, and it was nearly impossible to decide on a generally accepted value. As a result, we have decided to use 10-fold cross-validation on the training samples; in every single experiment, we chose C value by using the accuracy maximizing C value in 10-fold cross-validation.

4.3.2 Feature definition and selection

It is not possible to reach high classification rates without the use of expressive features. It was required to define the best selection of features for the success of constrained Delaunay triangulation, so we tried to define features on CDT that consider and cover the spatial relationships, connectivity, and network related properties of nodes.

To this end, we first made use of available features which are defined for Delaunay triangulation in the previous works of other researchers. However, these features were simply not sufficient to represent the characteristics of our constrained Delaunay triangulation, so we extended our feature set using the features defined for other graph types which are also available in the literature. Features like the number of isolated and end nodes, and the number of connected components are not meaningful if they are defined on a regular Delaunay triangulation, but they are so on constrained Delaunay triangulation. The combined set of these features are still not enough, so to uncover the real discriminative power of CDT, we have also defined features like average clustering coefficient for nodes with $d(p_i) \geq 2$. These features are summarized in Table 4.18.

Another problem was to select the most representative features out of this feature set. We made use of principle component analysis, forward selection, and backward selection to understand the relationships and the effects of features in a deeper manner. For the forward selection, we have added features until the classification rate obtained from 10-fold cross-validation increases no more, and for the backward selection, we continued to drop features until the same accuracy rate does not increase. However, when we examine the results presented in the corresponding sections, the results show that these algorithms could not improve the classification and increase accuracy. The problem with these algorithms is that once a feature is selected to be used in or removed from the feature subset, this feature cannot be removed or included in a later iteration.

Considering all these results, we have decided that the final settings for the algorithm should be as follows:

Feature	**DT**[1]	**OG**[2]	**CDT**[3]
Average degree	✓	✓	✓
Average degree for nodes with $d(p_i) \geq 2$		✓	✓
Isolated node number		✓	✓
End node number		✓	✓
Average clustering coefficient for nodes with $d(p_i) \geq 2$		✓	✓
Average eccentricity	✓	✓	✓
Diameter	✓	✓	✓
Number of components		✓	✓
Number of components with $n \geq 2$		✓	✓
Giant component ratio		✓	✓
Average edge length	✓	✓	✓
Standard deviation of edge length	✓	✓	✓
Average triangle area	✓		✓
Standard deviation of triangle area	✓		✓
[1] Features that are definable for Delaunay triangulation [2] Features that are definable for other types of graphs [3] Features that are definable for constrained Delaunay triangulation			

Table 4.18: The list of features

- K-means clustering with $k = 3$
- Preprocessing
- Circle transformation with a threshold value of 10
- Construction of the constrained Delaunay triangulation and the extraction of all features defined in Table 3.1
- Training and classification with SVM by using the cost parameter C found by 10-fold cross-validation, with the use of all features

With these settings, the obtained accuracy results and their comparison with those of Delaunay triangulation are given in Tables 4.19, 4.20, 4.21 and 4.22.

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	38	0	0	100.00
Low-grade	3	26	8	70.27
High-grade	0	3	37	92.50
Overall Accuracy				87.83

Table 4.19: Training set accuracy obtained by the constrained Delaunay triangulation.

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	31	0	3	91.18
Low-grade	2	25	8	71.43
High-grade	0	1	28	96.55
Overall Accuracy				85.71

Table 4.20: Test set accuracy obtained by the constrained Delaunay triangulation.

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	33	2	3	86.84
Low-grade	3	21	13	56.76
High-grade	2	3	35	87.50
Overall Accuracy				77.39

Table 4.21: Training set accuracy obtained by the Delaunay triangulation.

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	29	0	5	85.29
Low-grade	4	21	10	60.00
High-grade	1	3	25	86.21
Overall Accuracy				76.53

Table 4.22: Test set accuracy obtained by the Delaunay triangulation.

4.3.3 Complexity of algorithms

We should also analyze the computational complexity of our algorithms to demonstrate that the introduction of constrained Delaunay triangulation does not bring high computational complexity.

In the first group of algorithms in our methodology, the Lab conversion and the k-means processing is common for both Delaunay triangulation (DT) and constrained Delaunay triangulation (CDT), so there is no difference in complexity for DT and CDT. For preprocessing, circle-fit transform, and Delaunay triangulation steps, the white regions and purple regions are processed individually, and the computation time is usually doubled or quadrupled. However, the increase in the number of pixels/nodes for the algorithms in this group does not affect the asymptotic complexity.

<table>
<tr><th rowspan="2">Step</th><th colspan="2">Complexity</th></tr>
<tr><th>DT</th><th>CDT</th></tr>
<tr><td>Lab conversion</td><td colspan="2" rowspan="2">same</td></tr>
<tr><td>K-means clustering</td></tr>
<tr><td>Preprocessing</td><td colspan="2" rowspan="3">same</td></tr>
<tr><td>Circle-fit transform</td></tr>
<tr><td>Delaunay triangulation</td></tr>
<tr><td>Constrained Delaunay triangulation</td><td>-</td><td>$O(N)$</td></tr>
<tr><td>Feature extraction</td><td colspan="2" rowspan="2">same</td></tr>
<tr><td>Training and classification</td></tr>
</table>

Table 4.23: Complexity of algorithms

The formation of constrained Delaunay triangulation is simply $O(N)$ after the construction of Delaunay triangulation, where N is the total number of white and purple nodes. The decrease in the number of edges does not affect the complexity, but shortens the feature extraction step and decreases the run-time. The feature extraction, training, and classification algorithms also do not increase the computational complexity. In short, it can be concluded that these two approaches present the same computational complexity.

Chapter 5

Conclusion and Future Work

The increasing risk of cancer in the 21st century raises more challenges for a pathologist. The spread of contaminants and hormone-injected food, the rise in cigarette and alcohol consumption, stressful lifestyle, and many other factors cause more cancer incidents throughout the mankind, and this fact compels cancer specialists to give more accurate decision in shorter times. Given the variability of human decision, it becomes inevitable to employ computerized decision makers to help pathologists since algorithmic approaches offer more stable and quantitative frameworks.

There exists a large set of studies for automated cancer diagnosis, especially based on textural and/or structural tissue analysis. The major drawback of the previous structural, graph-based studies is their incapability of using potential information that is provided by other tissue components rather than cell nuclei. Because of their nature, such information becomes useful especially for the representation of the tissue types where tissues consist of hierarchical structures, such as gland structures in colon tissues.

In this thesis, we proposed a novel constrained Delaunay triangulation(CDT)-based technique for diagnosis and grading of colon cancer. For the construction of CDT, histopathological images of colon tissues are first transformed into Lab color space and the pixels of these images are clustered using the k-means clustering

algorithm. Resulting image maps are then preprocessed and circle-fit transform is applied on these ones, to define circular primitives which will represent nuclei and luminal regions. Afterwards, a standard Delaunay triangulation is built on these two sets of nodes (nuclei and luminal nodes). The constrained Delaunay triangulation is obtained with the removal of luminal (white) nodes and the edges that are connected to these nodes.

Our proposed algorithm utilizes luminal regions for the construction of triangulation. With the introduction of CDT, it becomes possible to come up with a new and distinctive set of features, which could not be defined on standard Delaunay triangulation (DT), such as number of isolated nodes, number of end nodes, and giant component ratio. These novel features are better in terms of reflecting the layout of a tissue with components rather than nuclei. On the other hand, features which are already available for Delaunay triangulation become more significant with the construction of a more expressive triangulation. Tables 5.1 and 5.2 demonstrate the training and test accuracies for constrained Delaunay triangulation, respectively. In these results, the features that are also available to standard Delaunay triangulation, such as average degree and average edge length, are used. The results show that the use of former features result in a classification with lower accuracies. The test set accuracy of 76.53 percent is 9.18 percent lower than the results that are acquired with the use of all features.

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	35	1	2	92.11
Low-grade	3	25	9	67.57
High-grade	0	4	36	90.00
Overall Accuracy				83.48

Table 5.1: Training set accuracy obtained by the constrained Delaunay triangulation. In this experiment, the common features that are also available to standard Delaunay triangulation are used for the training and classification.

After the acquisition of the features, classification is conducted with the use of a support vector machine classifier. 10-fold cross-validation is applied on the cross validation set, which basically is the training set, for the detection of optimal

	Healthy	**Low-grade**	**High-grade**	**Accuracy**
Healthy	25	3	6	73.53
Low-grade	2	23	10	65.71
High-grade	1	1	27	93.10
Overall Accuracy				76.53

Table 5.2: Test set accuracy obtained by the constrained Delaunay triangulation. In this experiment, the common features that are also available to standard Delaunay triangulation are used for the training and classification.

cost parameter C of SVM classifier.

In this study, principle component analysis, forward selection, and backward elimination methods are used to find an optimal subset of features. These methods provided the necessary information for our analysis of individual features, but they could not find the best feature subset. In forward selection algorithm, a selected feature cannot be removed at later iterations, even if it becomes useless with the insertion of other features. Furthermore, in backward elimination approach, a removed feature cannot be reincluded in the feature subset in backward elimination approach. Therefore, we have decided to use all available features in the training and classification steps.

The results indicate that, CDT provides better performance over DT in terms of classification accuracy. 85.71% accuracy for the test set is achieved with a classification over the use of whole feature set, while the Delaunay triangulation provides 76.53% for the same configuration. Our preselected feature subset provides better classification for the test set with 86.73% accuracy, but not for the training set. We choose to use the entire feature set, considering the lower accuracy acquired for the training set with the use of manually selected features, because the former produced more stable and non-random results. Considering their peak values, a gain of 10% is obtained without the introduction of any computational complexity.

The contribution of this thesis to literature is that, it proves the need for the utilization of non-nuclei components of various tissue types, in which these

components come into question, such as colorectal tissues. A novel constrained Delaunay triangulation is designed for the use of multi-component image analysis studies.

The future aspects of this study include running the experiments on other types of tissues with dominating structures. A second possible improvement and a future work shall be combining this idea of the utilization of non-nuclei components with the other structural approaches.

Bibliography

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

[2] F. Albregtsen, B. Nielsen, and H. E. Danielsen. Adaptive gray level run length features from class distance matrices. *Proceedings of the 15th International Conference on Pattern Recognition, ICPR*, 3:738–741, 2000.

[3] A. Andrion, C. Magnani, P. Betta, A. Donna, F. Mollo, M. Scelsi, P. Bernardi, M. Botta, and B. Terracini. Malignant mesothelioma of the pleura: interobserver variability. *Journal of Clinical Pathology*, 48(9):856–860, 1995.

[4] M. V. Anglada. An improved incremental algorithm for constructing restricted Delaunay triangulations. *Computers and Graphics*, 21(2):215–223, 1997. Graphics Hardware.

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

[6] F. Aurenhammer. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405, 1991.

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

[8] 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.

[9] G. Bigras, R. Marcelpoil, E. Brambilla, and G. Brugal. Cellular sociology applied to neuroendocrine tumors of the lung: Quantitative model of neoplastic architecture. *Cytometry*, 24(1):74–82, 1996.

[10] 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.

[11] Y. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, 1977.

[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] J. Bouttier, P. Di Francesco, and E. Guitter. Geodesic distance in planar graphs. *Nuclear Physics B*, 663:535, 2003.

[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] L. P. Chew. Constrained Delaunay triangulations. In *Proceedings of the Third Symposium on Computational Geometry*, pages 215–222, New York, NY, USA, 1987. ACM.

[17] L. P. Chew. Constrained Delaunay triangulations. *Algorithmica*, 4:97–108, 1989.

[18] 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.

[19] 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.

[20] P. Cignoni, C. Montani, and R. Scopigno. Dewall: A fast divide and conquer Delaunay triangulation algorithm in E^d. *Computer-Aided Design*, 30(5):333–341, 1998.

[21] R. W. Conners and C. A. Harlow. Theoretical comparison of texture algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(3):204–222, 1980.

[22] 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.

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

[24] P. C. Cosman, R. M. Gray, and M. Vetterli. Vector quantization of image subbands: A Survey. *IEEE Transactions on Image Processing*, 5(2):202–225, 1996.

[25] 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.

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

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

[28] L. de Floriani and E. Puppo. Constrained Delaunay triangulation for multiresolution surface description. *9th International Conference on Pattern Recognition*, 1:566–569, 1988.

[29] 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.

[30] B. N. Delaunay. Sur la sphère vide. *Bulletin of Academy of Sciences of the USSR*, 7:793–800, 1934.

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

[32] 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.

[33] G. Deng, I. Bell, S. Crawley, J. Gum, J. P. Terdiman, B. A. Allen, B. Truta, M. H. Sleisenger, and Y. S. Kim. Braf mutation is frequently present in sporadic colorectal cancer with methylated hmlh1, but not in hereditary nonpolyposis colorectal cancer. *Clinical Cancer Research*, 10(1):191–195, 2004.

[34] O. Devillers. Improved incremental randomized Delaunay triangulation. In *SCG '98: Proceedings of the Fourteenth Annual Symposium on Computational Geometry*, pages 106–115, New York, NY, USA, 1998. ACM.

[35] 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.

[36] 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.

[37] 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.

[38] E. Domingo, P. Laiho, M. Ollikainen, M. Pinto, L. Wang, A. J. French, J. Westra, T. Frebourg, E. Espin, M. Armengol, R. Hamelin, H. Yamamoto, R. M. W. Hofstra, R. Seruca, A. Lindblom, P. Peltomaki, S. N. Thibodeau, L. A. Aaltonen, and S. Jr. Schwartz. Braf screening as a low-cost effective strategy for simplifying HNPCC genetic testing. *Journal of Medical Genetics*, 41(9):664–668, 2004.

[39] V. Domiter and B. Zalik. Sweep-line algorithm for constrained Delaunay triangulation. *International Journal of Geographical Information Science*, 22(4):449–462, 2008.

[40] 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.

[41] R. A. Dwyer. A faster divide-and-conquer algorithm for constructing Delaunay triangulations. *Algorithmica*, 2:137–151, 1987.

[42] A. N. Esgiar, R. N. Naguib, M. K. Bennett, and A. Murray. Automated feature extraction and identification of colon carcinoma. *Analytical and Quantitative Cytology and Histology*, 20(4):297–301, 1998.

[43] 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.

[44] 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.

[45] R. Fabbri, L. F. Estrozi, and L. D. F. Costa. On Voronoi diagrams and medial axes. *Journal of Mathematical Imaging and Vision*, 17(1):27–40, 2002.

[46] L. P. Fielding, P. A. Arsenault, P. H. Chapuis, O. Dent, B. Gathright, J. D. Hardcastle, P. Hermanek, J. R. Jass, and R. C. Newland. Clinicopathological staging for colorectal cancer: An international documentation system (IDS) and an international comprehensive anatomical terminology (ICAT). *Journal of Gastroenterology and Hepatology*, 6(4):325–344, 1991.

[47] 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.

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

[49] S. Fortune. A sweepline algorithm for Voronoi diagrams. *Algorithmica*, 2:153–174, 1987.

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

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

[52] 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.

[53] 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.

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

[55] 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.

[56] J. Gudmundsson, H. J. Haverkort, and M. van Kreveld. Constrained higher order Delaunay triangulations. *Computational Geometry: Theory and Applications*, 30(3):271–277, 2005.

[57] L. Guibas, D. Knuth, and M. Sharir. Randomized incremental construction of Delaunay and Voronoi diagrams. *Algorithmica*, 7(1-6):381–413, 1992.

[58] L. Guibas and J. Stolfi. Primitives for the manipulation of general subdivisions and the computation of Voronoi. *ACM Transactions on Graphics*, 4(2):74–123, 1985.

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

[60] 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.

[61] 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.

[62] J. C. Hardwick. Implementation and evaluation of an efficient parallel Delaunay triangulation algorithm. In *SPAA '97: Proceedings of the Ninth Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 239–248, New York, NY, USA, 1997. ACM.

[63] 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.

[64] Z. Hu, H. Yan, and X. Lin. Clothing segmentation using foreground and background estimation based on the constrained Delaunay triangulation. *Pattern Recognition*, 41(5):1581–1592, 2008.

[65] S. Huerta. Recent advances in the molecular diagnosis and prognosis of colorectal cancer. *Expert Review of Molecular Diagnostics*, 8:277–288, 2008.

[66] 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.

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

[68] T. Joachims. Making large-scale SVM learning practical. *Advances in Kernel Methods: Support Vector Learning*, pages 169–184, 1999.

[69] 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.

[70] M. Kallmann, H. Bieri, and D. Thalmann. Fully dynamic constrained Delaunay triangulations. In G. Brunnett, B. Hamann, H. Mueller, and L. Linsen, editors, *Geometric Modelling for Scientific Visualization*, pages 241–257. Springer-Verlag, Heidelberg, Germany, first edition, 2003.

[71] M. Kandemir. Segmentation of colon glands by object graphs. Master's thesis, Bilkent University, Ankara, Turkey, 2008.

[72] 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.

[73] I. Kolingerová and B. Zalik. Improvements to randomized incremental Delaunay insertion. *Computers and Graphics*, 26(3):477–490, 2002.

[74] D. T. Lee and A. Lin. Generalized Delaunay triangulation for planar graphs. *Discrete and Computational Geometry*, 1:201–217, 1986.

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

[76] W. Lixin, W. Yanbing, and S. Wenzhong. Integral ear elimination and virtual point-based updating algorithms for constrained Delaunay TIN. *Science in China Series E: Technological Sciences*, 51(Supplement1):135–144, 2008.

[77] Y. Lu and W. W. Dai. A numerical stable algorithm for constructing constrained Delaunay triangulation and application to multichip module layout. *International Conference on Circuits and Systems, 1991. Conference Proceedings, China*, 2:644–647, 1991.

[78] H. T. Lynch, J. F. Lynch, and P. M. Lynch. Toward a consensus in molecular diagnosis of hereditary nonpolyposis colorectal cancer (lynch syndrome). *Journal of the National Cancer Institute*, 99(4):261–263, 2007.

[79] 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.

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

[81] R. Marcelpoil and Y. Usson. Methods for the study of cellular sociology: Voronoi diagrams and parametrization of the spatial relationships. *Journal of Theoretical Biology*, 154(4):359–369, 1992.

[82] K. Masood, N. Rajpoot, K. Rajpoot, and H. Qureshi. Hyperspectral colon tissue classification using morphological analysis. *International Conference on Emerging Technologies, 2006*, pages 735–741, 2006.

[83] 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.

[84] E. F. Moore. The shortest path through a maze. In *Proceedings of the International Symposium on the Theory of Switching*, pages 285–292. Harvard University Press, 1959.

[85] P. Morrison and J. J. Zou. Triangle refinement in a constrained Delaunay triangulation skeleton. *Pattern Recognition*, 40(10):2754–2765, 2007.

[86] J. Muckell, M. Andrade, W. R. Franklin, B. Cutler, M. Inanc, Z. Xie, and D. Tracy. Hydrology-aware constrained triangulation of terrain data. (unpublished), 2008.

[87] J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, and R. Arbiol. Multiresolution-based image fusion with additive wavelet decomposition. *IEEE Transactions on Geoscience and Remote Sensing*, 37(3):1204–1211, 1999.

[88] 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.

[89] D. Park, H. Cho, and Y. Kim. A TIN compression method using Delaunay triangulation. *International Journal of Geographical Information Science*, 15:255–269, 2001.

[90] W. Qian, L. P. Clarke, B. Zheng, M. Kallergi, and R. Clark. Computer assisted diagnosis for digital mammography. *IEEE Engineering in Medicine and Biology Magazine*, 14(5):561–569, 1995.

[91] K. Rajpoot and N. Rajpoot. SVM optimization for hyperspectral colon tissue cell classification. In *Proceedings of MICCAI 2004*, pages 829–836, 2004.

[92] K. Rajpoot, N. Rajpoot, and M. Turner. Hyperspectral colon tissue cell classification. In *Proceedings of SPIE Medical Imaging*, volume 1, pages 1–2, SPIE, 2004.

[93] 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.

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

[95] Y. L. C. e. a. Robert F Ozols, Roy S Herbst. Clinical cancer advances 2006: Major research advances in cancer treatment, prevention, and screening – A report from the american society of clinical oncology. *Journal of Clinical Oncology*, 25(1):146–162, 2007.

[96] D. L. Rubbelke. *Tissues of the human body: An introduction.* McGraw-Hill, 1999.

[97] J. Ruppert. A new and simple algorithm for quality 2-dimensional mesh generation. In *SODA '93: Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 83–92, Philadelphia, PA, USA, 1993. Society for Industrial and Applied Mathematics.

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

[99] 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.

[100] D. Satyanarayana and S. V. Rao. Constrained Delaunay triangulation for ad hoc networks. *Journal of Computer Systems, Networks, and Communications*, 2008(2):1–10, 2008.

[101] G. Schaller and M. Meyer-Hermann. Multicellular tumor spheroid in an off-lattice Voronoi-Delaunay cell model. *Physical Review E*, 71(5), 2005.

[102] J. Scholefield, H. Abcarian, A. Grothey, and T. Maughan, editors. *Challenges in Colorectal Cancer, 2nd Edition.* John Wiley and Sons, 2^{nd} edition, 2006.

[103] O. Sertel, J. Kong, H. Shimada, U. Catalyurek, J. Saltz, and M. Gurcan. Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. *Pattern Recognition*, In Press, Corrected Proof, 2008.

[104] V. B. Shenoy, R. Miller, E. B. Tadmor, D. Rodney, R. Phillips, and M. Ortiz. An adaptive finite element approach to atomic-scale mechanics – the quasicontinuum method. *Journal of the Mechanics and Physics of Solids*, 47(3):611–642, 1999.

[105] J. R. Shewchuk. A condition guaranteeing the existence of higher-dimensional constrained Delaunay triangulations. In *SCG '98: Proceedings of the Fourteenth Annual Symposium on Computational Geometry*, pages 76–85, New York, NY, USA, 1998. ACM.

[106] J. R. Shewchuk. Sweep algorithms for constructing higher-dimensional constrained Delaunay triangulations. In *SCG '00: Proceedings of the Sixteenth Annual Symposium on Computational Geometry*, pages 350–359, New York, NY, USA, 2000. ACM.

[107] 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.

[108] 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.

[109] R. I. Silveira and M. van Kreveld. Towards a definition of higher order constrained Delaunay triangulations. *Computational Geometry*, 42(4):322–337, 2009.

[110] M. C. Southey, M. A. Jenkins, L. Mead, J. Whitty, M. Trivett, A. A. Tesoriero, L. D. Smith, K. Jennings, G. Grubb, S. G. Royce, M. D. Walsh, M. A. Barker, J. P. Young, J. R. Jass, D. J. B. St John, F. A. Macrae, G. G. Giles, and J. L. Hopper. Use of molecular tumor characteristics to prioritize mismatch repair gene testing in early-onset colorectal cancer. *Journal of Clinical Oncology*, 23(27):6524–6532, 2005.

[111] 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.

[112] 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.

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

[114] 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.

[115] 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.

[116] 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.

[117] V. J. D. Tsai. Delaunay triangulations in TIN creation: an overview and a linear-time algorithm. *International Journal of Geographical Information Science*, 7(6):501–524, 1993.

[118] A. Üngör. Off-centers: A new type of steiner points for computing size-optimal quality-guaranteed Delaunay triangulations. *Computational Geometry*, 42(2):109–118, 2009.

[119] M. Vigo and N. Pla. Computing directional constrained Delaunay triangulations. *Computers and Graphics*, 24(2):181–190, 2000.

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

[121] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[122] J. Weszka, C. R. Dyer, and A. Rosenfeld. A comparative study of texture measures for terrain classification. *IEEE Transaction on Systems, Man and Cybernetics*, 6(4):269–285, 1976.

[123] 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.

[124] 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.

[125] 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.

[126] 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.

[127] WHO. Cancer. Technical Report 297, World Health Organization, WHO Media Centre, 2008.

[128] Y. Xue, M. Sun, and A. Ma. On the reconstruction of three-dimensional complex geological objects using Delaunay triangulation. *Future Generation Computer Systems*, 20(7):1227–1234, 2004. Geocomputation.

[129] W. Yanbing, W. Lixin, and S. Wenzhong. Constrained edge dynamic deleting in CD-TIN based on influence domain retriangulation of virtual point. *Geo-spatial Information Science*, 10(3):208–212, 2007.

[130] B. Zalik. An efficient sweep-line Delaunay triangulation algorithm. *Computer-Aided Design*, 37(10):1027–1038, 2005.

[131] B. Zalik and I. Kolingerová. An incremental construction algorithm for Delaunay triangulation using the nearest-point paradigm. *International Journal of Geographical Information Science*, 17:119–138, 1 March 2003.

Appendix A

Implementation

Step	**Programming language**
Transformation into Lab color space	MATLAB
K-means clustering	ANSI C
Preprocessing	MATLAB
Circle-fit transform	ANSI C
Delaunay triangulation	MATLAB
Constrained Delaunay triangulation	MATLAB
Feature extraction	Java
Training and classification	ANSI C
PCA	MATLAB
Forward selection	MATLAB
Backward elimination	MATLAB

Table A.1: Implementation details of our approach