# AUTOMATIC IDENTIFICATION OF ABNORMAL REGIONS IN DIGITIZED HISTOLOGY CROSS-SECTIONS OF COLONIC TISSUES AND ADENOCARCINOMAS USING QUASI-SUPERVISED LEARNING

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in Electronics and Communication Engineering

by
Devrim ÖNDER

July 2012
İZMİR

We approve the thesis of **Devrim ÖNDER**

**Examining Committee Members:**

**Assoc. Prof. Dr. Bilge KARAÇALI**
Department of Electrical and Electronics Engineering
İzmir Institute of Technology

**Prof. Dr. Sülen SARIOĞLU**
Department of Pathology, Faculty of Medicine, Dokuz Eylül University

**Assist. Prof. Dr. Şevket GÜMÜŞTEKİN**
Department of Electrical and Electronics Engineering
İzmir Institute of Technology

**Assoc. Prof. Dr. Mehmet ENGİN**
Department of Electrical and Electronics Engineering, Ege University

**Assist. Prof. Dr. Zübeyir ÜNLÜ**
Department of Electrical and Electronics Engineering
İzmir Institute of Technology

**5 July 2012**

**Assoc. Prof. Dr. Bilge KARAÇALI**
Supervisor, Department of Electrical and Electronics Engineering
İzmir Institute of Technology

**Prof. Dr. Acar SAVACI**
Head of the Department of
Electrical and Electronics Engineering

**Prof. Dr. R. Tuğrul SENGER**
Dean of the Graduate School
Engineering and Sciences

# ACKNOWLEDGEMENTS

I would like to express my gratitude to the thesis advisor Assoc. Prof. Dr. Bilge Karaçalı not only for his guidance but also for his support, trust and recommendations throughout my thesis study. He also made the study easier since he was the main actor in setting-up the computer infrastructure in our laboratory and administrating the network. I would also send my friendly thanks to him for serving delicious elegant coffee in our review meetings.

I would like to express my gratitude to my committee members, Prof. Dr. Sülen Sarıoğlu and Assist. Prof. Dr. Şevket Gümüştekin for their contributions and guidance. Prof. Dr. Sülen Sarıoğlu enabled us performing inter-disciplinary research between engineering and histopathology. Thanks for her enthusiasm and tireless expert efforts.

I am also grateful to Biomedical Information Processing Laboratory (Biplab) members Başak Esin Köktürk and Tunca Doğan for their share and long technical discussions. They were always friendly to me.

I especially would like to thank Devrim Tokcan for his unfailing encouragement, never-ending friendship and support. Besides, I sincerely thank to all who guided, encouraged and supported me before the doctoral program.

Finally, I thank my parents Ercan Önder, Halide Önder and my brother Canderim Önder for their endless encouragement and loving support during my whole life.

# ABSTRACT

## AUTOMATIC IDENTIFICATION OF ABNORMAL REGIONS IN DIGITIZED HISTOLOGY CROSS-SECTIONS OF COLONIC TISSUES AND ADENOCARCINOMAS USING QUASI-SUPERVISED LEARNING

In this thesis, a framework for quasi-supervised histopathology image texture identification is presented. The process begins with extraction of texture features followed by a quasi-supervised analysis. Throughout this study, light microscopic images of the hematoxylin and eosin stained colorectal histopathology sections containing adenocarcinoma were quantitatively analysed. The quasi-supervised learning algorithm operates on two datasets, one containing samples of normal tissues labelled only indirectly and in bulk, and the other containing an unlabelled collection of samples of both normal and cancer tissues. As such, the algorithm eliminates the need for manually labelled samples of normal and cancer tissues commonly used for conventional supervised learning and significantly reduces the expert intervention. Several texture feature vector datasets corresponding to various feature calculation parameters were tested within the proposed framework. The resulting labelling and recognition performances were compared to that of a conventional powerful supervised classifier using manually labelled ground-truth data that was withheld from the quasi-supervised learning algorithm. That supervised classifier represented an idealized but undesired method due to extensive expert labelling. Several vector dimensionality reduction techniques were evaluated an improvement in the performance. Among the alternatives, the Independent Component Analysis procedure increased the performance of the proposed framework. Experimental results on colorectal histopathology slides showed that the regions containing cancer tissue can be identified accurately without using manually labelled ground-truth datasets in a quasi-supervised strategy.

# ÖZET

## KOLOREKTAL DOKU VE KOLOREKTAL ADENOKARSİNOM SAYISAL HİSTOLOJİ KESİTLERİNDEKİ ANORMAL BÖLGELERİN YARI-GÜDÜMLÜ ÖĞRENME KULLANILARAK OTOMATİK BELİRLENMESİ

Bu çalışmada histopatolojik görüntülerdeki desenlerin yarı-güdümlü öğrenme yardımıyla tanınması amacıyla bir metodoloji geliştirilmiştir. Sözkonusu metodoloji, desen vektörlerinin hesaplanması ve sonrasında uygulanan yarı-güdümlü öğrenme aşamalarından oluşmaktadır. Bu çalışma kapsamında, adenokarsinom içeren hematoksilen ve eosin boyama uygulanmış kolorektal histopatoloji kesitleri analize tabi tutulmuştur. Yarı-güdümlü öğrenme yöntemi, biri dolaylı yolla sağlıklı olarak belirlenmiş, diğeri üzerinde hiç bir işaretleme yapılmamış sağlıklı/kanserli örnekler olmak üzere iki ayrı vektör grubu üzerinde çalışır. Bu sayede, yarı-güdümlü öğrenme yöntemi geleneksel güdümlü öğrenme yöntemlerinin ihtiyaç duyduğu sağlıklı ve kanserli bölgelerin teker teker işaretlemesine gerek duymaz ve dolayısıyla uzman müdahelesi ihtiyacını önemli ölçüde azaltır. Önerilen yöntem, çeşitli desen vektör hesaplama parametreleri kullanılarak elde edilmiş desen vektör veri kümeleri üzerinde kullanılmıştır. Önerilen yöntemin desen tanıma başarımları, tümü elle işaretlenmiş desenleri kullanan kabul görmüş bir güdümlü öğrenme yöntemi ile karşılaştırılmıştır. Sözkonusu güdümlü öğrenme yöntemi fazla uzman müdahelesi gerektirdiği için alternatif sunduğumuz güdümlü yöntemlerinden birisidir. Bu çalışmada önerilen yöntemin tanıma başarımını arttırmak amacıyla çeşitli vektör boyut azaltma yaklaşımları denenmiştir. Bu yaklaşımlardan Bağımsız Bileşenler Analizi yöntemi önerilen yönteme ait desen tanıma başarımını arttırmıştır. Sonuç olarak, kolorektal histopatoloji kesitleri üzerinde yapılan deneyler, kanserli bölgeler içeren doku kesitlerinin tek tek elle işaretlenmiş verileri kullanmadan başarılı bir şekilde tanındığını göstermiştir.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| CAD | Computer-assisted Diagnosis |
| CPVMVH | Corresponding Pixel Value for Maximum Value of Histogram |
| CRCa | Colorectal Carcinoma |
| HIAE | Histopathological Image Atlas Editor |
| HIL | Histopathological Image Library |
| H&E | Hematoxylin and Eosin |
| ICA | Independent Component Analysis |
| IMC1 | Information Measures of Correlation 1 |
| IMC2 | Information Measures of Correlation 2 |
| MDS | Multidimensional Scaling |
| NNCR | Non-neoplastic Colorectal |
| PCA | Principal Component Analysis |
| ROC | Receiver Operating Characteristics |
| SVM | Support Vector Machine |
| QSL | Quasi-supervised Learning |

# CHAPTER 1

# INTRODUCTION

Improvements in image analysis and machine learning techniques allowed researchers to address the ultimate goal of supporting pathologists in disease detection and grading. Increasing number of complex patterns that need to be checked by pathologists and rapidly growing histopathology slide databases, keep the subject of automated quantitative analysis of histopathology slides indispensable today. Computerized analysis of histopathology slides has been a very attractive research topic with the recent advances in computational power.

Histopathology refers to the microscopic examination of tissue in order to study the manifestations of disease. Specifically, in clinical medicine, histopathology refers to the examination of a biopsy or surgical specimen by a pathologist, after the specimen has been processed and histological sections have been placed onto glass slides. The diagnosis from a histopathology image still remains the standard in diagnosing considerable number of diseases including almost all types of cancer (Rubin et al. 2011).

Recent developments in histopathology increased the importance of digital storage and processing of tissue slides in computerized environments. The development of full-automated image analysis systems that scan and segment normal/abnormal tissue profiles in recorded digital histology slides also became an appealing topic.

Cancer is a disease that can easily be identified by abnormal tissue profiles. Generally, heterogeneous cancerous regions can easily be identified in homogeneously distributed tissue profiles. Colorectal cancer is one of the commonest malignant tumors worldwide and represented the fifth cause of cancer-related death in 2008 (World Health Organization 2008). In Turkey, colorectal cancer was the fourth cause of cancer-related death due to 2008 statistics. Practically, colon or rectum cancer is characterized as separate cancer instances. Colorectal cancer is a composite name for colon and rectum cancer. It is the uncontrolled growth of tissue cells in either the colon or rectum which causes the colorectal cancer. The vast majority of colorectal cancer is an adenocarcinoma (nearly 85%) because the colon has numerous glands within the tissue (Lanza et al. 2011).

In a supervised pattern classification algorithm, which is usually a part of the automated disease detection, an expert should individually check and label all the training features into several categories in order to have a ground truth dataset. Manual processing effort of the biomedical data might be very time consuming especially for today's rapidly growing image databases. In addition, objective decisions held for some features may have a negative effect on the performance of the supervised classifier used. Especially, for multi-user platforms or for a system that is implemented in a distributed architecture, suppression of the expert objectivity becomes widely important. In computational analysis of histology slides, for instance, a pathologist can easily identify tissue cross-sections that are free from cancerous abnormalities. However, such abnormalities may occur amid tissue that is benign in appearance and have to be either painstakingly labelled by an expert pathologist.

The main objective of our study is to perform automated detection of colorectal adenocarcinomas in a set of light microscopic histopathology slide images by minimising the prior expert intervention by a quasi-supervised learning algorithm. Quasi-supervised Learning (QSL) is a statistical learning algorithm that contrasts two datasets by computing estimates for the posterior probabilities of their samples of belonging in either dataset (Karaçalı 2010). Therefore, the QSL method is suitable for a cancer disease identification problem whether the labelled samples would have been available from one class only, which is healthy, and, a second, unlabelled dataset could also be provided containing a mixture of samples from both reference and target classes, which are healthy and cancer, respectively.

Hence, we have constructed a histopathology image dataset including images of healthy tissues only, defined as non-neoplastic colorectal (NNCR), and images partially containing CRCa tissues. We also organised the histopathology image dataset as two separate histopathology image groups; the reference group containing only healthy images and the unlabelled mixed group having either the healthy images or images having CRCa tissues partially. From the perspective of feature class labels, we had two data groups, the reference group including features of class label NNCR and the mixed group including features of both class labels, either NNCR and CRCa.

QSL algorithm is to be applied to the reference and mixed feature data groups and the disease identification is to be carried out by contrasting the unlabelled mixed dataset samples to the reference dataset. QSL selects those that are dissimilar from the reference samples beyond a statistical significance level as target samples.

Vector dimensionality reduction is a mathematical transformation to represent a vector dataset in a relatively lower dimension. In a classification problem, the dimensionality reduction is usually performed to improve the classification performance and the computation time.

In this study, to detect heterogeneous target texture profiles, multi-dimensional textural feature vectors are to be calculated using the local histogram based first order texture features and features derived from co-occurrence matrices which are known as Haralick features (Haralick et al. 1973).

This study is basically focused on the evaluation of QSL on various feature extraction parameter configurations and different dimensionality reduction approaches that is applied to the original data. It is also very important to gather the comparative information of a new method against other off the shelf vector classifiers on the same texture feature vector datasets in order to comment about the performance as an alternative approach. In order to obtain an independent evaluation of the labelling and target classification performances, we have used a Support Vector Machine (SVM) classifier trained on the ground-truth label data that was withheld from the quasi-supervised labelling strategy (Cortes and Vapnik 1995, Vapnik 1998, Burges 1998).

## 1.1 Organization Of The Thesis

This dissertation is organized as follows; Chapter 2 is dedicated to the description of the problem with details on histopathology science and disease to be evaluated (i.e. colorectal adenocarcinoma). A literature review will follow in Chapter 3 summarizing the research previously done on quantitative histopathological image analysis and quasi-supervised learning. In Chapter 4, the technical background information for the elementary parts of the proposed framework is presented. In addition, the proposed framework is described in the form of a graphical abstract. The experimental setup of our study along with the results of the experimental execution is described in Chapter 5. Furthermore, Chapter 6 consists of the conclusions of the thesis research and a discussion about future projections that may follow this research effort.

# CHAPTER 2

# PROBLEM DESCRIPTION

This chapter basically describes our study. The main idea of the study was to perform automated detection of a specific disease by minimising the need for prior expert intervention, by a quasi-supervised learning algorithm. The problem definition could be specified more by selecting light microscopy histopathology imaging as modality and selecting a specific disease; human colorectal adenocarcinoma (CRCa) which originates in colon tissues.

To have a mature problem definition, we should first introduce the histopathology science and histology of the colon. In addition, colorectal cancers which have increasing incidence and mortality rates in recent years, are to be described below.

## 2.1 Histopathology

Histopathology (compound of three Greek words: histos "tissue", pathos "disease-suffering", and logia) refers to the microscopic examination of tissue in order to study the manifestations of disease. Specifically, in clinical medicine, histopathology refers to the examination of a biopsy or surgical specimen by a pathologist, after the specimen has been processed and histological sections have been placed onto glass slides. In contrast, cytopathology examines free cells or tissue fragments.

Histopathological examination of tissues starts with surgery, biopsy or autopsy. The tissue is removed from the body or plant, and then placed in a fixative which stabilizes the tissues to prevent decay. The most common fixative is formalin that is 10% formaldehyde in water.

The tissue is then prepared for viewing under a microscope using either chemical fixation or frozen section which will be discussed in the following sections.

## Chemical Fixation

In chemical fixation, the samples are transferred to a cassette, a container designed to allow reagents to freely act on the tissue inside. This cassette is immersed in multiple baths of progressively more concentrated ethanol, to dehydrate the tissue, followed by toluene or xylene and finally extremely hot liquid (usually paraffin). During this 12 to 16 hour process, paraffin will replace the water in the tissue, turning soft, moist tissues into a sample miscible with paraffin, a type of wax. This process is known as tissue processing.

The processed tissue is then taken out of the cassette and set in a mold. Through this process of embedding, additional paraffin is added to create a paraffin block which is attached to the outside of the cassette.

The process of embedding then allows the sectioning of tissues into very thin (2 - 7 micrometer) sections using a microtome. The microtome slices the tissue ready for microscopic examination. The slices are thinner than the average cell, and are layered on a glass slide for staining.

## Frozen Section Processing

The second method of histopathology processing is called frozen section processing. In this method, the tissue is frozen and sliced thinly using a microtome mounted in a below-freezing refrigeration device called the cryostat. The thin frozen sections are mounted on a glass slide, fixed immediately and briefly in liquid fixative, and stained using the similar staining techniques as traditional wax embedded sections. The advantages of this method is rapid processing time, less equipment requirement, and less need for ventilation in the laboratory. The disadvantage is the poor quality of the final slide. It is used in intra-operative pathology for determinations that might help in choosing the next step in surgery during that surgical session (e.g., to preliminarily determine clearness of the resection margin of a tumor during surgery).

## Staining of the Processed Histopathology Slides

In order to see the tissue under a microscope, the sections are stained with one or more pigments. The aim of staining is to reveal cellular components; counterstains are used to provide contrast. The majority of stains only absorb light, and the stained slides are therefore viewed using a microscope with a light illuminating the sample from below. If no stain is present, all of the light will pass through, appearing bright white. Areas where the stain has adhered to a substance in the tissue will absorb some of the light. The amount of light absorbed depends on many factors. For a given unit of stain, a certain amount of light in each spectrum will be absorbed.

The most commonly used stain in histopathology is a combination of hematoxylin and eosin (often abbreviated H&E). Hematoxylin is used to stain nuclei blue, while eosin stains cytoplasm and the extracellular connective tissue matrix pink. Due to the long history of H&E, well-established methods, and a tremendous amount of data and publications, there is a strong belief among many pathologists that H&E will continue to be the common practice over the next 50 years (Fox 2000).

There are hundreds of various other techniques which have been used to selectively stain cells. Other compounds used to color tissue sections include safranin, Oil Red O, congo red, silver salts and artificial dyes.

## Interpretation

The histological slides are examined under a microscope by a pathologist, a medically qualified specialist who has completed a recognised training programme. This medical diagnosis is formulated as a pathology report describing the histological findings and the opinion of the pathologist. In the case of cancer, this represents the tissue diagnosis required for most treatment protocols. In the removal of cancer, the pathologist will indicate whether the surgical margin is cleared, or is involved (residual cancer is left behind).

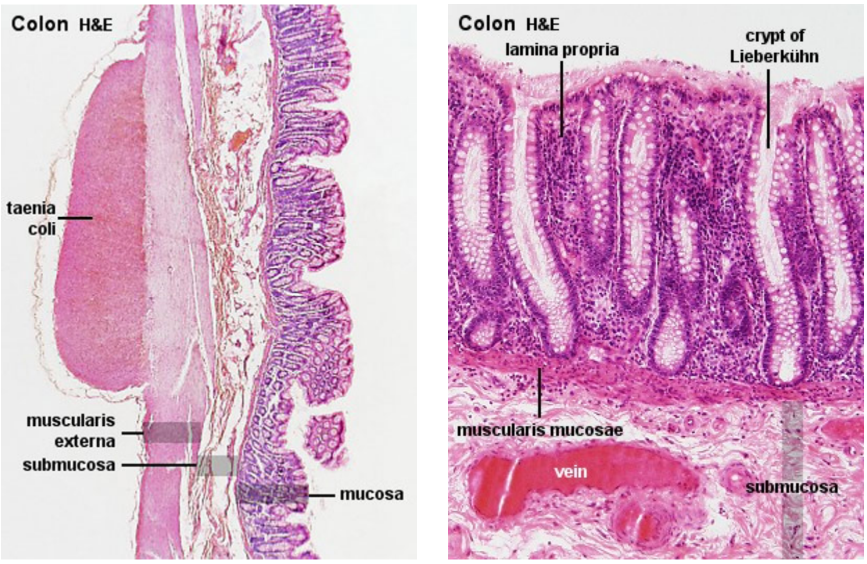
## 2.2 Histology Of The Colon

The colon is the upper part of the large intestine tube while the rectum is the lower part. The colon is the last part of the digestive system in most vertebrates; it extracts water and salt from solid wastes before they are eliminated from the body, and is the site in which flora-aided (largely bacterial) fermentation of unabsorbed material occurs. Unlike the small intestine, the colon does not play a major role in absorption of foods and nutrients. However, the colon does absorb water, sodium and some fat soluble vitamins.

Basic histological parts of the colon include;

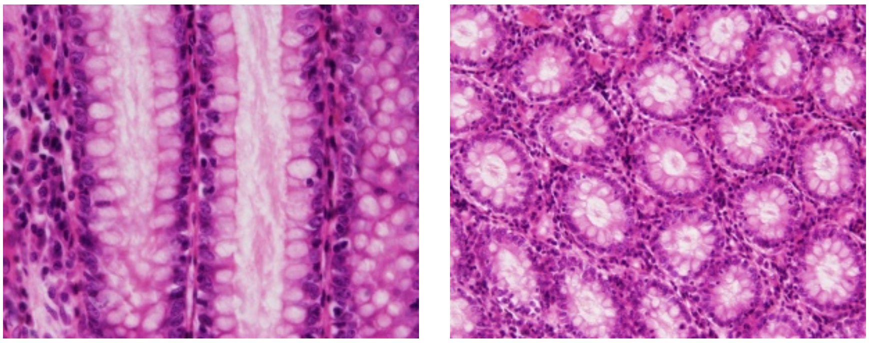
- Mucosa within which exists the epithelium, the intestinal glands (glands of Lieberkühn), lamina propria and muscularis mucosa.
    - lamina propria - (lamina propria mucosae), the layer of loose connective tissue beneath the gastrointestinal tract epithelium and with the epithelium form the mucosa.
    - muscularis mucosa - thin layer of smooth muscle outside the lamina propria and separating it from the submucosa of the gastrointestinal tract, this layer ends at the recto-anal junction.
- Submucosa is dense irregular connective tissue that supports the mucosa.
- Muscularis externa; containing inner circular and outer longitudinal smooth muscle layers.
- Lymphatic nodules in the lamina propria and submucosa.
- A myenteric (Auerbach) nerve plexus (parasympathetic) exists between the muscularis externa layers.
- The outermost serosa which is the outermost connective tissue layer covering the gastrointestinal tract in regions where it passes through body cavities.

Figures 2.1 - 2.6 illustrate the histological parts on several H & E stained colon sections.



(a) (b)

Figure 2.1. Colon histology.
(Source: UNSW Embryology website.)



(a) (b)

Figure 2.2. Crypts of Lieberkühn - (intestinal gland, intestinal crypt) (a) longitudinal, (b) transverse. (Source: UNSW Embryology website.)

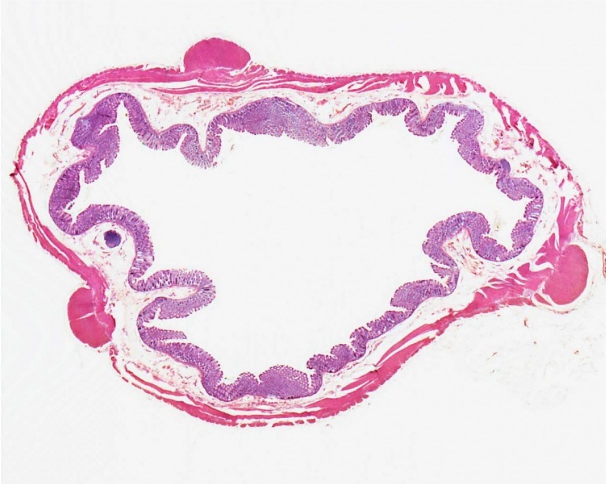


Figure 2.3. Colon cross section histological view.
(Source: Notes On Gastrointestinal Histology, University of Ottawa)

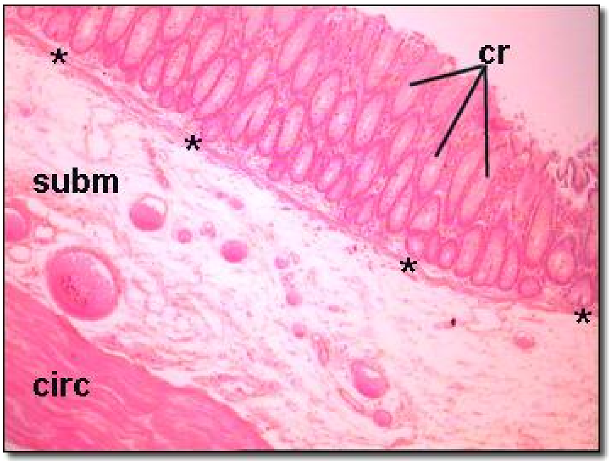


Figure 2.4. Low power magnification view of the colon with glands cut obliquely. (cr) crypts or glands, (*) muscularis mucosae, (subm) submucosa, (circ) inner circular layer of muscularis externa. (Source: Notes On Gastrointestinal Histology, University of Ottawa )

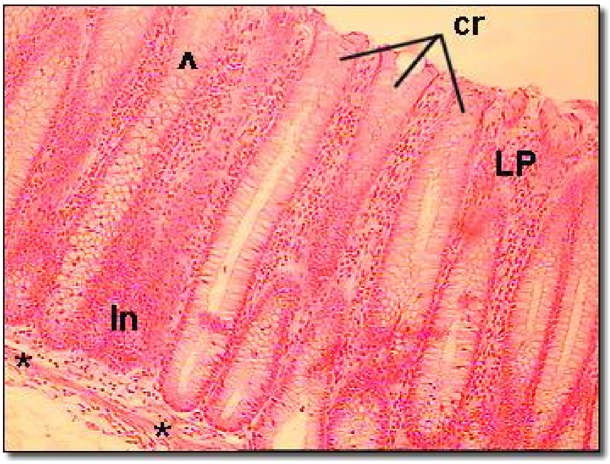


Figure 2.5. Higher power magnification view of mucosa of colon. (cr) crypts or glands, (LP) lamina propria, (ln) lymph nodule, (*) muscularis mucosae, inside one gland, (∧) shows change in sectioning from straight (top) to more oblique. (Source: Notes On Gastrointestinal Histology, University of Ottawa )

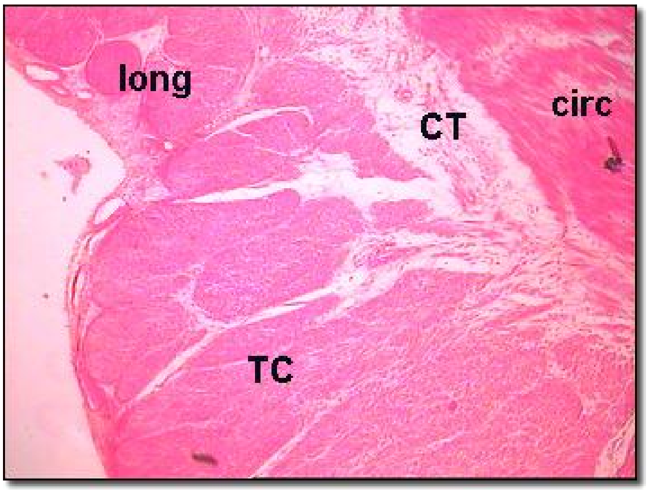


Figure 2.6. Low power magnification view of muscularis externa with tenia coli. (circ) inner circular layer, (CT) connective tissue separating circular and longitudinal muscle layer, (long) outer longitudinal layer, (TC) part of a tenia coli. (Source: Notes On Gastrointestinal Histology, University of Ottawa )

## 2.3 Adenocarcinoma

Adenocarcinoma is a cancer of an epithelium that originates in glandular tissue. Epithelial tissue includes, but is not limited to, the surface layer of skin, glands and a variety of other tissue that lines the cavities and organs of the body. Epithelium can be derived embryologically from ectoderm, endoderm or mesoderm. To be classified as adenocarcinoma, the cells do not necessarily need to be part of a gland, as long as they have secretory properties. This form of carcinoma can occur in some higher mammals, including humans. Well differentiated adenocarcinomas tend to resemble the glandular tissue that they are derived from, while poorly differentiated adenocarcinomas may not. By staining the cells from a biopsy, a pathologist can determine whether the tumor is an adenocarcinoma or some other type of cancer. Adenocarcinomas can arise in many tissues of the body due to the ubiquitous nature of glands within the body. While each gland may not be secreting the same substance, as long as there is an exocrine function to the cell, it is considered glandular and its malignant form is therefore named adenocarcinoma. Endocrine gland tumors, such as a VIPoma, an insulinoma, a pheochromocytoma, etc., are typically not referred to as adenocarcinomas, but rather, are often called neuroendocrine tumors. If the glandular tissue is abnormal, but benign, it is said to be an adenoma. Benign adenomas typically do not invade other tissue and rarely metastasize. Malignant adenocarcinomas invade other tissues and often metastasize given enough time to do so.

### 2.3.1 Colorectal Adenocarcinoma

Practically, colon or rectum cancer is characterized as separate cancer instances. Colorectal or bowel cancer is a composite name for colon and rectum cancer. It is the uncontrolled growth of tissue cells in either the colon or rectum which causes the colorectal cancer.

Epithelial tumors of the colon and rectum are frequent pathologic entities. Colorectal cancer is one of the commonest malignant tumors worldwide and represented the fifth cause of cancer-related death in 2008, please see (World Health Organization 2008). In order to see the estimated worldwide incidence and mortality rates please refer to Figures 2.7 and 2.8.

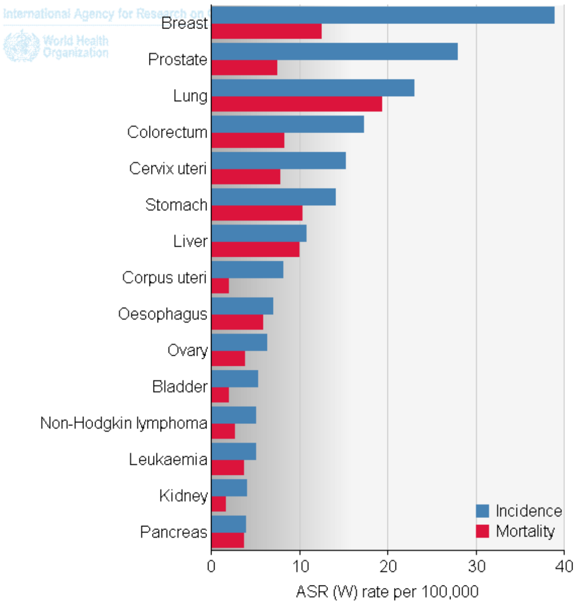
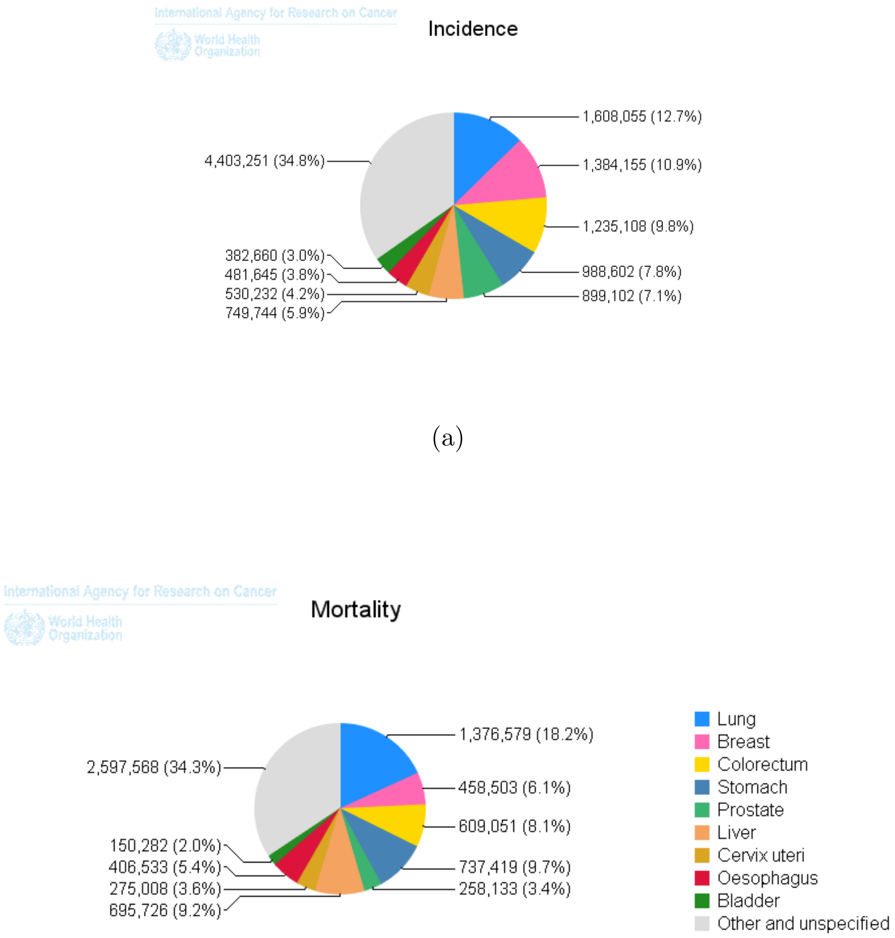


Figure 2.7. Estimated age-standardised incidence and mortality rates, for both sexes: 2008, Worldwide. (Source: World Health Organization)

In Turkey, colorectal cancer was the fourth cause of cancer-related death due to 2008 statistics. Figures 2.9 and 2.10 present the corresponding estimated incidence and mortality rates.

The vast majority of colorectal cancer is an adenocarcinoma (nearly 85%) (Lanza et al. 2011). This is because the colon has numerous glands within the tissue. Normal colonic glands tend to be simple and tubular in appearance with a mixture of mucus secreting goblet cells and water absorbing cells. These structures are called glands because they secrete a substance into the lumen of the colon, this substance being mucus. The purpose of these glands are twofold. The first is to absorb water from the feces back into the blood. The second purpose is to secrete mucus into the colon lumen to lubricate the now dehydrated feces. This is crucial as a failure to lubricate the feces can result in colonic damage by the feces as it passes towards the rectum.

When these glands undergo a number of changes at the genetic level, they



(a)

(b)

Figure 2.8. Estimated age-standardised (a) incidence and (b) mortality rates, for both sexes: 2008, Worldwide. (Source: World Health Organization)

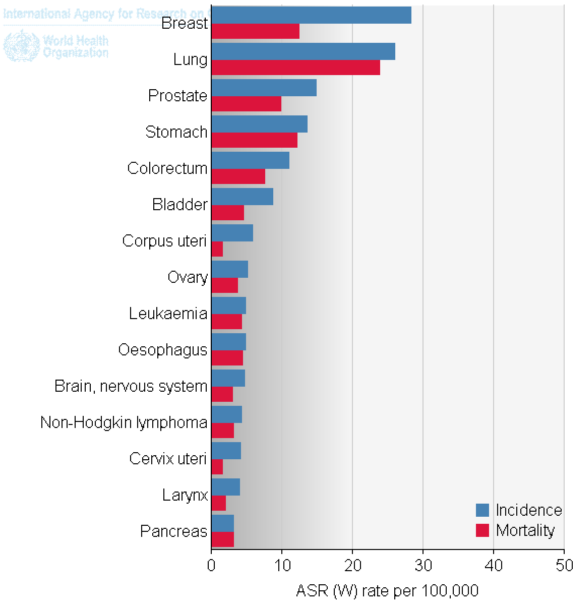
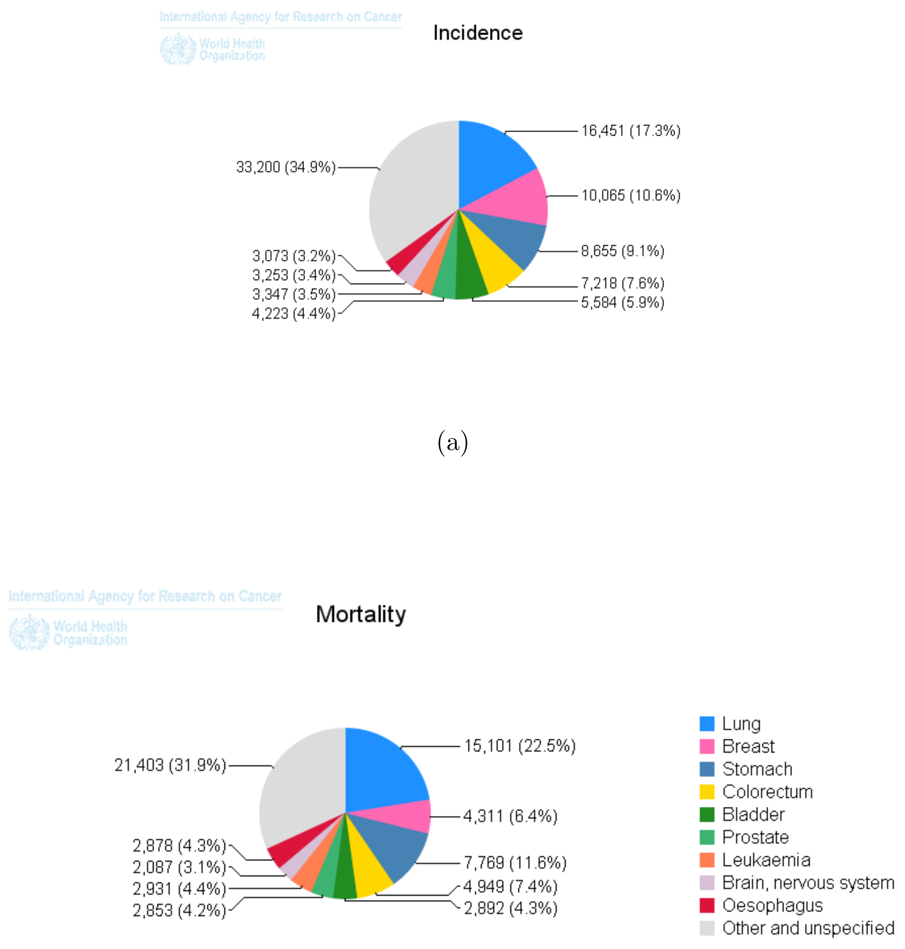


Figure 2.9. Estimated age-standardised incidence and mortality rates, for both sexes: 2008, Turkey. (Source: World Health Organization)

proceed in a predictable manner as they move from benign to an invasive, malignant colon cancer. In their research paper "Lessons from Hereditary Colorectal Cancer", Vogelstein, et al., suggested that colon cells lose the APC tumor suppressor gene and become a small polyp (Kinzler and Vogelstein 1996). Next, they suggested that k-Ras gene becomes activated and the polyp becomes a small, benign, adenoma. The adenoma, lacking the "carcinoma" attached to the end of it, suggests that it is a benign version of the malignant adenocarcinoma. The gastroenterologist uses a colonoscopy to find and remove these adenomas and polyps to prevent them from continuing to acquire genetic changes that will lead to an invasive adenocarcinoma. Volgelstein et al. went on to suggest that loss of the DCC gene and of p53 tumor suppressor protein result in a malignant adenocarcinoma.

Grossly, one will see a mass that looks of a different color than the surrounding tissue. Bleeding from the tumor is often apparent as the tumor tends to grow blood vessels into it in a haphazard manner via secretion of a number of angiogenesis promoting factors such as VEGF. Histologically, tumor resembling original structures are classified as well differentiated. Tumor cells, that have lost any resemblance to original tissue, both in appearance and structure form are denoted as poor differentiated tumor cells. Regardless of the grade, malignant tumors tend to have



(a)

(b)

Figure 2.10. Estimated age-standardised (a) incidence and (b) mortality rates, for both sexes: 2008, Turkey. (Source: World Health Organization)

a large nucleus with prominent nucleoli. There will also be a noticeable increase in the incidence of mitoses, or cell divisions.

Colorectal adenocarcinoma is histologically characterised with any one of the following conditions;

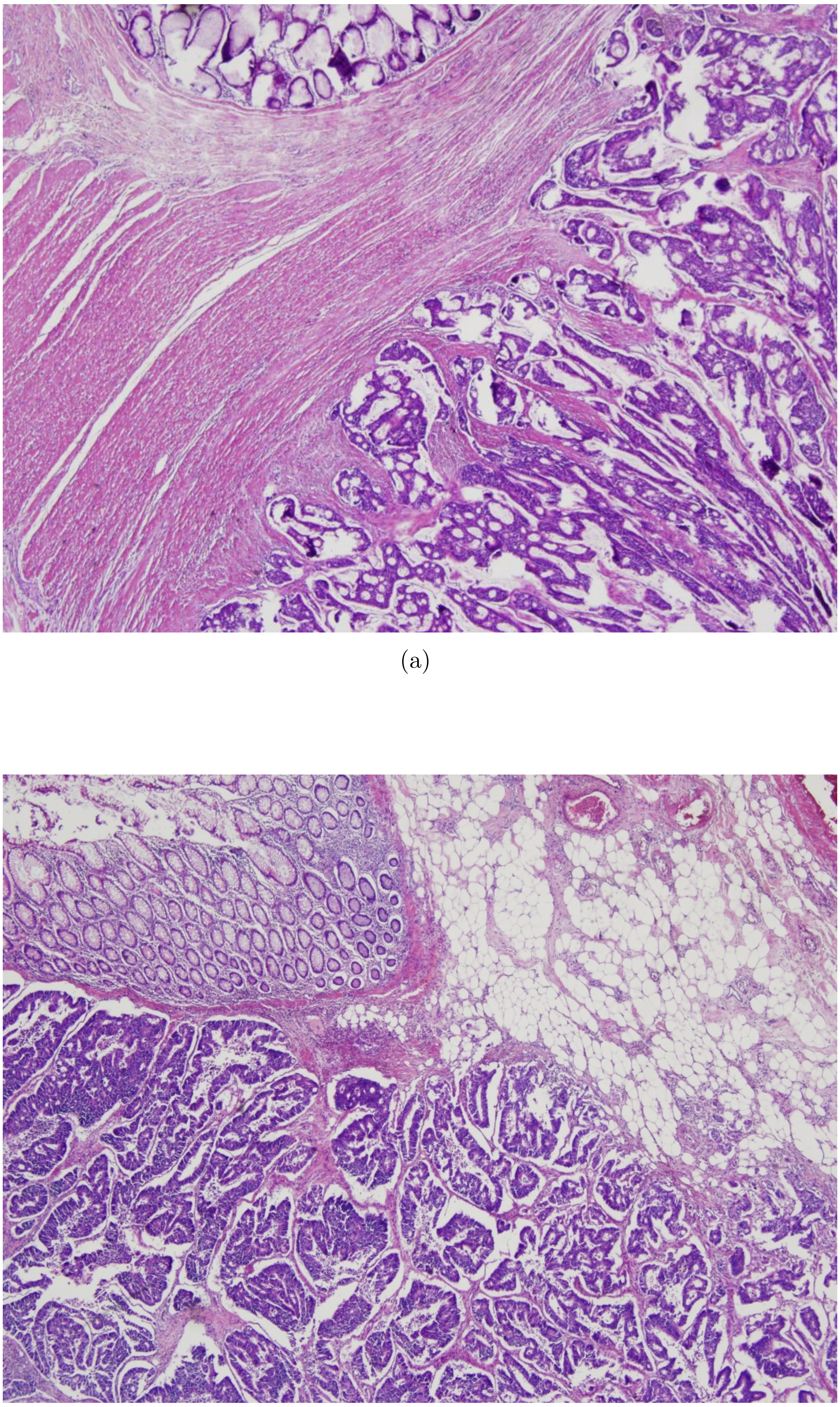
- Usually significant architectural and/or cytologic atypia is/are observed.
- Desmoplastic stroma exists.
- Glands not accompanied by lamina propria.
- Tumor invades submucosa.
- Tumor invades muscularis propria.
- Tumor invades through the muscularis propria into the subserosal adipose tissue or the nonperitonealized pericolic or perirectal soft tissues.
- Tumor penetrates to the surface of the visceral peritoneum (serosa).
- Tumor is usually infiltrative, non-circumscribed.

In Figure 2.11, two histopathology slides containing colorectal adenocarcinoma are presented. In addition, there are severeal future reference histopathlogy slides for colorectal adenocarcinoma are available in Figures 5.1 , 5.14(b) and 5.20.

### 2.3.2 Overview Of Colorectal Adenocarcinoma Diagnosis Methods

There are several different methods available in order to diagnose colorectal cancer, please see (Society 2012) and (Medical 2012) . These methods are discussed below;

- Blood tests: Colorectal adenocarcinoma can be diagnosed or even previously diagnosed cases can be monitored by certain blood tests. These blood tests include;



(a)

(b)

Figure 2.11. Two histopathology slides containing colorectal adenocarcinoma.

  - Fecal occult blood test (FOBT): FOBT is a test for blood in the stool. Two types of tests are used for detecting occult blood in stools i.e. guaiac based (chemical test) and immunochemical. The sensitivity of immunochemical testing is superior to that of chemical testing without an unacceptable reduction in specifity.
  - Complete blood count (CBC): CBC is performed to detect anemia (too few red blood cells). Some people with colorectal cancer become anemic because of prolonged bleeding from the tumor.
  - Liver enzymes: Since colorectal cancer can spread to the liver, a blood test is performed to check the liver functions.

- Tests to look for colorectal cancer or colorectal polyps:
  - Digital rectal exam (DRE): The doctor inserts a lubricated, gloved finger into the rectum to feel for abnormal areas. It only detects tumors large enough to be felt in the distal part of the rectum but is useful as an initial screening test.
  - Endoscopy. If symptoms or the results of the physical exam or blood tests suggest that colorectal cancer might be present, the doctor may recommend more tests. This most often is colonoscopy, but sometimes a sigmoidoscopy or an imaging test.
    - Sigmoidoscopy: A lighted probe (sigmoidoscope) is inserted into the rectum and lower colon to check for polyps and other abnormalities.
    - Colonoscopy: A lighted probe called a colonoscope is inserted into the rectum and the entire colon to look for polyps and other abnormalities that may be caused by cancer. Colonoscopy is a good option because it allows doctors to inspect the entire length of the colon with a little camera. Colonoscopy has the advantage that if polyps are found during the procedure they can be immediately removed. Tissue can also be taken for biopsy.
  - Biopsy: Usually if a suspected colorectal cancer is found by any diagnostic test, it is biopsied during a colonoscopy. In a biopsy, the doctor removes a

small piece of tissue with a special instrument passed through the scope. Less often, part of the colon may need to be surgically removed to make the diagnosis. No colon cancer diagnosis is final until the tissue sample from the colon is analyzed in a laboratory and found to contain cancer cells.

  - Lab tests of samples: Biopsy samples (from colonoscopy or surgery) are sent to the lab where a pathologist, a doctor trained to diagnose cancer and other diseases in tissue samples, looks at them under a microscope. Other tests may suggest that colorectal cancer is present, the only way to determine this for certain is to look at the samples under a microscope.
  - Stool DNA testing is an emerging technology in screening for colorectal cancer. Premalignant adenomas and cancers shed DNA markers from their cells which are not degraded during the digestive process and remain stable in the stool. Capture, followed by Polymerase Chain Reaction (PCR) amplifies the DNA to detectable levels for assay. Clinical studies have shown a cancer detection sensitivity of 71%-91%.

- Imaging tests by using different modalities: Imaging tests use sound waves, x-rays, magnetic fields, or radioactive substances to create pictures of the inside of the human body. Imaging tests may be done for a number of reasons, including to help find out whether a suspicious area might be cancerous, to learn how far cancer may have spread, and to help determine if treatment has been effective.

  - Computed tomography (CT or CAT) scan. The CT scan is an x-ray test that produces detailed cross-sectional images of your body. Instead of taking one picture, like a regular x-ray, a CT scanner takes many pictures as it rotates around a patient who is lying on a table. A computer then combines these pictures into images of slices of the part of your body being studied. Unlike a regular x-ray, a CT scan creates detailed images of the soft tissues in the body. This test can help tell if colon cancer has spread into your liver or other organs.
  - Ultrasound. Ultrasound uses sound waves and their echoes to produce a

picture of internal organs or masses. A small microphone-like instrument called a transducer emits sound waves and picks up the echoes as they bounce off body tissues. The echoes are converted by a computer into a black and white image that is displayed on a computer screen. This test does not expose any radiation to the patient.

Abdominal ultrasound can be used to look for tumors in your liver, gallbladder, pancreas, or elsewhere in your abdomen, but it can't look for tumors of the colon. For the exam, you simply lie on a table and a technician moves the transducer along the skin overlying the part of your body being examined.

Two special types of ultrasound exams are sometimes used to evaluate colon and rectal cancers:

  - Endorectal ultrasound: This test uses a special transducer that is inserted directly into the rectum. It is used to see how far through the rectal wall a cancer may have penetrated and whether it has spread to nearby organs or tissues such as lymph nodes.
  - Intraoperative ultrasound: This exam is done during surgery after the surgeon has opened the abdominal cavity. The transducer can be placed against the surface of the liver, making this test very useful for detecting the spread of colorectal cancer to the liver.

- Magnetic resonance imaging (MRI) scan. Like CT scans, MRI scans provide detailed images of soft tissues in the body. But MRI scans use radio waves and strong magnets instead of x-rays. The energy from the radio waves is absorbed by the body and then released in a pattern formed by the type of body tissue and by certain diseases. A computer translates the pattern into a very detailed image of parts of the body. A contrast material called gadolinium is often injected into a vein before the scan to better see details.

  MRI scans are sometimes useful in looking at abnormal areas in the liver that might be due to cancer spread. They can also help determine if rectal cancers have spread into nearby structures. To improve the accuracy of the test, some doctors use endorectal MRI. For this test the doctor places a probe, called an endorectal coil, inside the rectum.

- Positron emission tomography (PET) scan. For a PET scan, a form of radioactive sugar (known as fluorodeoxyglucose or FDG) is injected into the blood. The amount of radioactivity used is very low. Cancer cells in the body grow rapidly, so they absorb large amounts of the radioactive sugar. After about an hour, the patient will be moved onto a table in the PET scanner. The patient lies on the table for about 30 minutes while a special camera creates a picture of areas of radioactivity in the body. The picture is not finely detailed like a CT or MRI scan, but it provides helpful information about the whole body.

  A PET scan can help give the doctor a better idea of whether an abnormal area seen on another imaging test is a tumor or not. If one has already been diagnosed with cancer, the doctor may use this test to see if the cancer has spread to lymph nodes or other parts of the body.

- X-ray double contrast barium enema (DCBE): In this method, firstly, an overnight preparation is taken to cleanse the colon. An enema containing barium sulfate is administered, then air is insufflated into the colon, distending it. The result is a thin layer of barium over the inner lining of the colon which is visible on x-ray films. A cancer or a precancerous polyp can be detected this way. This technique can miss the (less common) flat polyps.

## 2.4 Objectives Of The Study

The main objective of the study was to perform automated detection of colorectal adenocarcinomas in a set of light microscopic histopathology slide images by minimising the prior expert intervention by QSL algorithm.

In a supervised pattern classification algorithm, which is usually a part of the automated disease detection, an expert should individually check and label all the training features into several categories in order to have a ground truth dataset. Ground truth datasets are then used as the basis for statistical learning, specifically to construct a classification rule using one of the methods; support vector machines (Cortes and Vapnik 1995, Vapnik 1998), nearest neighbour classifiers (Cover and Hart 1967), neural networks (Haykin 2008), discriminant functions (McLachlan 2004).

Manual processing effort of the biomedical data might be very time consuming especially for today's rapidly growing image databases. In addition, objective decisions held for some features may have a negative effect on the performance of the supervised classifier used. Especially, for multi-user platforms or for a system that can be implemented in a distributed architecture, suppression of the expert objectivity becomes widely important. In computational analysis of histology slides, for instance, a pathologist can easily identify tissue cross-sections that are free from cancerous abnormalities. However, such abnormalities may occur amid tissue that is benign in appearance and have to be either painstakingly labelled by an expert pathologist.

Quasi-supervised Learning (QSL) is a statistical learning algorithm that contrasts two datasets by computing estimates for the posterior probabilities of their samples of belonging in either dataset (Karaçalı 2010). Therefore, the QSL method is suitable for a our disease identification problem whether the labelled samples would have been available from one class only, which is healthy. A second, unlabelled dataset could also be provided containing a mixture of samples from both reference and target classes, which are healthy and cancer, respectively. Hence, we should set up a histopathology image dataset including images of healthy tissues only and images including CRCa tissues. We will also organise the histopathology image dataset as two separate histopathology image groups; the reference group containing only healthy images and the unlabelled mixed group having either the healthy images or images having CRCa tissues partially. From the perspective of feature class labels, we will have two data groups, the reference group including features of class label NNCR and the the mixed group including features of both class labels NNCR and CRCa.

QSL algorithm is to be applied to the reference and mixed feature groups and the disease identification is to be carried out by contrasting the unlabelled mixed dataset samples to the reference dataset, and selects those that are dissimilar from the reference samples beyond a statistical significance level as target samples.

In this study, we decided to use local image texture characteristics other than object level identifiers. We plan to perform tests with various texture characteristics and to determine which characteristics will be the most discriminative for the proposed methodology. Especially, in cases where the data do not allow perfect separation of the both classes, the algorithm is expected to outperform off-the-shelf classification algorithms as it is a model-free alternative to existing techniques and it will not search for a separation boundary optimized according to some criteria. In order to have performance comparison, a supervised classifier will be applied to the same feature datasets and the resultant classification results will be reviewed.

On the other hand, several dimensionality reduction algorithms used in automated pattern classification schemes will also be tested in our methodology to capture if there is a better data characterisation and related performance improvement in accurate texture identification.

# CHAPTER 3

# LITERATURE REVIEW

This chapter presents an overview of the existing literature on the fields of automated diagnosis and quantitative histopathological image analysis. The entry part makes an introduction with the history of the computer-assisted diagnosis (CAD) researches. In the first section, a specialised adoption of CAD, the quantitative histopathological image analysis, is reviewed. In the review of quantitative histopathological image analysis studies, firstly, several studies are described by focusing corresponding tissues of interest. In its following subsections, the studies are presented in a separate subsection per each main functional stages of a histopathological image analysis framework (i.e color normalisation, segmentation, feature extraction and classification). In the later section, resultant classification accuracies reported in several image analysis studies are listed. These accuracy levels given in percentage terms enable readers to compare with our output labelling and classification performances. In the last section, which describes the literature on “Quasi-supervised Learning”, we present several recent researches on QSL paradigm and along with their outputs.

Computer-assisted diagnosis research can be traced back to the 1980s (Diamond et al. 1982, Hallouche et al. 1992). Furthermore, the widespread use of CAD has started with the emergence of digital mammography in the early 1990s (Sahiner et al. 1996, Mendez et al. 1998). Currently, CAD has become one of the major research subjects in medical imaging and diagnostic radiology.

With the recent advances in digitized histological archives caused by high-throughput tissue banks, it now became possible to use histological tissue patterns with computer-aided image analysis to perform disease classification. There is also an expanding need for CAD to relieve the workload on pathologists by detecting obviously benign areas, so that the pathologist can focus on the more suspicious cases. As an example, approximately 80% of prostate biopsies performed in the U.S. every year are benign; this suggests that prostate pathologists are spending most of their time examining benign tissue. (Gurcan et al. 2009)

A large focus of pathological image analysis has been on the automated analysis of cytology imagery, since, cytology imagery often results from the least invasive biopsies (e.g., the cervical Pap smear). Additionally, the characteristics of cytology imagery, namely the presence of isolated cells and cell clusters in the images and the absence of more complicated structures such as glands make it easier to analyse these specimens compared to histopathology.

On the other hand, histopathology slides, provide a more comprehensive view of a disease and its effects, since the underlying tissue architecture is preserved by the preparation process. Therefore, some disease characteristics, such as lymphocytic infiltration of cancer, may be deduced only from a histopathology image. In addition, the diagnosis from a histopathology image still remains the standard in diagnosing considerable number of diseases including almost all types of cancer (Rubin et al. 2011).

## 3.1 Quantitative Histopathological Image Analysis

This section presents an overview of the literature in the field of histopathological image analysis. The first part makes an introduction to histopathological image analysis researches held from the tissue of interest, disease to be diagnosed or to be graded point of view. In the following subsection, past researches on color normalisation, which is very important for the cases including unbalanced histopathology slide images are discussed. Following this subsection, various feature extraction procedures taking part in histopathological image analysis are described. The later subsection is the literature review of the studies grouped by classification paradigms, or using the general terminology, the machine learning algorithms. Subsection 3.2 is reserved for the review of colonic histopathological image analysis researches held, since we also deal with the same tissue type in our thesis.

Computerized analysis of histopathology slides has been a very attractive research topic with the recent advances in computational power. Meanwhile, improvements in image analysis and machine learning techniques allowed researchers to address the ultimate goal of supporting pathologists in diagnosis, disease detection and grading. Increasing number of complex patterns that need to be checked by pathologists and rapidly growing histopathology slide databases keep the subject of automated quantitative analysis of histopathology slides indispensable today.

There have been many histopathology studies conducted for automated detection of regions possessing the characteristics of a specific disease. The diseases taken into usual consideration in this respect include renal cell carcinoma (Waheed et al. 2007), breast (Sahiner et al. 1996), cervical cancer (Hallouche et al. 1992) and prostate cancer (Diamond et al. 1982, Pitts et al. 1993, Doyle et al. 2006, 2007). Researches performed investigating colon tissues are detailed in a dedicated section (see Section 3.2).

### 3.1.1 Color Normalisation In Histopathological Image Analysis

Inconsistencies in the preparation of histology slides make it difficult to perform quantitative analysis. Researchers provided various algorithms for overcoming many of the known inconsistencies in the staining process, thereby bringing slides that were processed or stored under very different conditions into a common, normalized space to enable improved quantitative analysis.

In many biological fields, the sections are stained with one or more pigments in order to see the tissue under a microscope. The aim of staining is to reveal cellular components; counter-stains are used to provide contrast (for more detail please see Section 2.1). The overall amount of light absorbed also varies between slides prepared differently. The two most prominent factors that affect the intensity of a slide are the relative amounts of stain added and the subsequent storage and handling of the slide, as stains can fade when exposed to light.

Most histology slides are examined in isolation by a pathologist. These examinations focus on relative color differences and morphology of biological features. Therefore, the pathologist need to compare different slides seldom. On the other hand, several software tools that perform correction for spectral and spatial illumination variations is becoming a standard package provided by most bright field manufacturers. This is an essential step for automated algorithms that heavily depend on color space computations. This process reduces the differences in tissue samples due to variation in staining and scanning conditions. The illumination can be corrected either using calibration targets or estimating the illumination pattern from a series of images by fitting polynomial surfaces (Can et al. 2008). Another approach is to match the histograms of the images.

A simple algorithm was presented to obtain stain saturation values when the stain vectors describing how the color is affected by the stain concentration (Ruifrok and Johnston 2001). The remaining issue was determining which stain vectors should be used. Previously calculated approximations for each type of stain exist but these approximations ignore variations between specific stains. The proposed method in the original paper requires manual selection of an area on the slide that contains only one stain, and then calculates an average stain vector from this area. In order to avoid contamination, it was recommended that a slide be stained with only one stain at a time. Obviously, individually staining a slide is not a viable option for slides that have already been processed or when only a scan of the slide is available. This leaves the inferior option of manually selecting an area with a minimal amount of other stains. Although this approach will lead to better results than using a pre-calculated approximation, it is very tedious for large datasets.

Rabinovich et. al. proposed non-negative matrix factorization (NMF) to solve the general color un-mixing problem (Rabinovich et al. 2004). This study was motivated by the high levels of user interaction needed to determine the stain vectors. Yang et. al. proposed another robust color-based segmentation algorithm for histological structures that used image gradients estimated in the *LUV* color space to deal with issues of stain variability (Yang et al. 2005). Wang et. al., in their oral cancer tissue classification scheme, used a color normalisation scheme proposed by Reinhard et. al. (Reinhard et al. 2001, Wang et al. 2007). The method used was a simple statistical analysis to impose one image's color characteristics on another. The researchers stated that one can achieve color correction by choosing an appropriate source image and applying its characteristic to another image.

Macenko et. al. performed a study of color normalisation in optical density space using H&E stained slides of melanomas and nevi (Macenko et al. 2009). The authors stated that the method was applicable to other histologic stains and tissues.

The algorithm for obtaining the optimal stain vectors has been evaluated on slides with various stain combinations satisfactorily. However, when three or more stains were present in a slide, results were sometimes found inconsistent.

In another color normalisation study, Magee et. al. presented a couple of colour normalisation algorithms for digital histology images, evaluated against linear normalisation in *Lab* colour space (Magee et al. 2009). These procedures mapped the colour distribution of an over/under stained image to that of a well stained target image. The first method was based on the linear normalisation in *Lab* colour space method, extended to multiple pixel classes using a probabilistic (Gaussian Mixture Model based) colour segmentation method. Linear normalisation was applied separately for each pixel class (where class membership is defined by a pixel being coloured by a particular chemical stain, or being uncoloured i.e. background). This approach assumed an additive colour model. The second method presented was based on normalisation in a stain specific colour deconvolution representation (Ruifrok and Johnston 2001). In this representation, each dimension represents the amount by which a pixel is stained by a particular chemical stain. There were two variants of the procedure; a linear normalisation was applied in this representation, and, separate transforms were defined for foreground and background. Magee et. al. concluded that the segmentation based approach, while producing good results on the majority of images, was less successful than the colour deconvolution method for a significant minority of images as robust segmentation was required to avoid introducing artefacts.

### 3.1.2 Histopathological Image Feature Extraction

This section lists the previous approaches to feature extraction in quantitative histopathological image analysis. In order to provide easy consideration, Gurcan et. al divided feature extraction methodologies into basic groups as; object level, multi-scale and spatially related (Gurcan et al. 2009). We here followed the same structure for the review of feature extraction.

## Object Level Feature Extraction

Commonly, an object in image analysis is defined as a connected group of pixels satisfying some similarity criteria. Object-level histopathological analysis depends greatly on some underlying segmentation mechanism. Despite the main focus is often on the segmentation of nuclei, there exists little work that explicitly using features of cytoplasm and stroma. Naik et al. used cytoplasmic and stromal features to automatically segment glands in prostate histopathology (Naik et al. 2007). In that study, the classification performance in distinguishing between different grades of prostate cancer was found to be comparable using manual and automated gland and nuclear segmentation. Moreover, it appeared that histologic objects may not need to be perfectly segmented to be properly classified when a list of comprehensive features was used in a feature selection scheme (Boucheron 2008). These results suggested that the perfect segmentation is not necessarily a prerequisite for a successful classification.

The definitions for object level features can be found in (Boucheron 2008). These features were compiled from a comprehensive literature search on cytopathology and histopathology image analysis. Object-level features can be categorized as belonging to one of four categories: size and shape, radiometric and densitometric, texture, and chromatin-specific.

In addition, various statistical measures for any of the vector quantities were also proposed. Thus, the statistical measures; mean, median, minimum, maximum, standard deviation, skewness, and kurtosis were calculated for all vector features.

Another approach that semantically describes histopathology images using model based intermediate representation (MBIR) and incorporates low-level color texture analysis was presented in (Sertel et al. 2009). In this approach, basic cytological components in an image were first identified using an unsupervised clustering in the *Lab* color space. The connected components of nuclei and cytoplasm regions were modelled using ellipses. An extensive set of features can be constructed from this intermediate representation to characterize the tissue morphology as well as tissue topology. Using this representation, the relative amount and spatial distribution of these cytological components can be measured. In the application of follicular lymphoma grading, where the spatial distribution of these regions varies considerably between different histological grades, MBIR provides a convenient way to quantify the corresponding observations.

In this thesis, the second order texture features were calculated by using co-occurrence matrices and the first order texture features were generated from local histograms either in gray level or *Lab* planes separately. The mathematical background and the implementation are described in detail in Section 4.1 and Section 5.3.

## Multi-scale Feature Extraction

Resulting from the density of the data and the fact that pathologists tend to employ a multi-resolution approach to analyse histopathology data, feature values are strongly related to the viewing scale or resolution. For instance, at low scales, color or texture cues are commonly used and at medium scales architectural arrangement of individual histological structures (glands and nuclei) start to become resolvable. It is only at higher resolutions that morphology of specific histological structures can be discerned.

In a couple of studies, a multi-resolution approach has been used for the classification of high-resolution whole-slide histopathology images (Kong et al. 2009, Sertel et al. 2009). The proposed multi-resolution approach mimics the evaluation of a pathologist such that image analysis starts from the lowest resolution, corresponding to the lower magnification levels in a microscope and uses the higher resolution representations for the regions requiring more detailed information for a classification decision. To achieve this, images were decomposed into multi-resolution representations using the Gaussian pyramid approach (Burt and Adelson 1983). This was followed by color space conversion and feature construction followed by feature extraction and feature selection at each resolution level. Once the classifier was confident enough at a particular resolution level, the system assigned a classification label (e.g., stroma-rich, stroma-poor or undifferentiated, poorly differentiating, differentiating) to the image tile. The resulting classification map from all image tiles forms the final classification map. The classification of a whole slide image is achieved by dividing into smaller image tiles and processing each image tile independently in parallel on a cluster of computer nodes.

In this thesis, a multi-scale feature extraction approach which we called "hierarchical texture feature vector calculation" was also evaluated. Our main approach was to bond all scale features together rather than decomposing and into separate resolution levels. For the mathematical background and specific in-house implementation, see Section 4.1.

### Spatially Related Feature Extraction

Graphs are also efficient data structures to represent spatial data and an effective way to represent structural information by defining a large set of topological features. The use of spatial-relation features for quantifying cellular arrangement was proposed in the early 1990s (Albert et al. 1992). Graphs have now been constructed for modelling different tissue states and to distinguish one state from another by computing metrics on these graphs and classifying their values. Overall, the use of spatial arrangement of histological entities (generally at low resolutions) is relatively new, especially in comparison to the wealth of research on nuclear features (at higher resolutions). Definitions for all graph structures and features can be found in (Boucheron 2008). The total number of spatial-relation features extracted was approximately 150 for all graph structures.

Doyle et. al. constructed the Voronoi diagram from a set of seed-like points that denote the centers of each structure of interest which is nuclei (Doyle et al. 2007). From the Voronoi diagram, two more graphs of interest can be constructed; the Delaunay triangulation, which is created by connecting points that share an edge in the Voronoi diagram, and the minimum spanning tree, which is the series of lines that spans the set of points such that the Euclidean sum of the lengths of the lines is smaller than any other spanning tree. From each of these three graphs, a series of features are calculated that captures the size, shape, and arrangement of the structures of the nuclei.

### 3.1.3 Classification In Histopathological Image Analysis

In histopathology image imagery, unlike some other applications of image analysis, one of the primary considerations in the choice of a classifier is its ability to deal with large, highly dense datasets.

Machine learning algorithms are basically divided into two groups as supervised and unsupervised classifiers. The common machine learning approach in almost all of the histopathology studies was the use of a supervised classification procedure to label the tissue regions of interest. To perform supervised classification studies, manually labelled training data were needed for the corresponding classifier training. In the following part, we have reviewed several histopathological image analysis articles indicating hosted supervised/unsupervised classification approaches.

In an example of unsupervised classification approach, Onder et. al. performed k-means clustering algorithm following a dimensionality reduction procedure, to discriminate basic texture characteristics in renal histopathology specimens (Önder and Karaçalı 2009).

As explained before, supervised machine learning algorithms were usually evaluated in histopathological image analysis. We now present a quick review of various supervised procedures and its applications;

Support vector machine (SVM) was one of the mostly used supervised classifier algorithms as reviewed in Section 4.3 in detail. We visited several histopathological image analysis articles based on SVM classifier; SVM was used to carry out discrimination between normal and malignant colon tissue cells (Nasir et al. 2004, Masood et al. 2006). Nasir et.al. and Masood et.al. used an SVM classifier, to carry out discrimination between normal and malignant colon tissue cells. SVM has also been used to differentiate colon adenocarcinoma histopathology images from benign ones (Rajpoot and Rajpoot 2004). A commonly used kernel known as the radial basis function has been employed to distinguish between three different prostate tissue classes in (Doyle et al. 2007). In another histopathological image analysis scheme, SVM has been also used to classify four different subtypes of meningioma (Qureshi et al. 2008). For more detailed background and evaluation of SVM please visit Section 4.3.

Among the classification strategies put to the task, both linear discriminate function and k-nearest-neighbour non-parametric classifiers were separately used to identify cancerous colonic mucosa (Esgiar et al. 1998b), automated classification of colorectal dysplasia, aspecific and ulcerative colitis were experimented by using discriminant analysis method (Hamilton et al. 1997, Ficsor et al. 2008).

Filippas et. al. focused on the identification of normal and cancerous colonic mucosa using a genetic algorithm (Filippas et al. 2003). Nwoye et. al. used a fuzzy neural network classifier to detect adenomas and adenocarcinomas in colorectal tissue slides (Nwoye et al. 2006). Doyle et. al. performed studies for automated prostatic adenocarcinoma detection and for prostate cancer grading using Adaboost, decision trees and SVM classifiers (Doyle et al. 2006, 2007). Computer-aided histopathological classification of renal cell carcinoma using a multi-class Bayesian decision rule that assumes multivariate Gaussian distributions for the feature vectors was also studied (Waheed et al. 2007).

## 3.2 Quantitative Histopathological Image Analysis In Colorectal Tissues

This section is particularly reserved for the comprehensive review of colonic histopathological image analysis researches held, since we analysed the same tissue pattern in our thesis study. The choice of discriminating features, the classifiers evaluated and the resultant classification accuracy values are presented together for each research.

One of the first attempts to computer-aided histopathological classification was carried out by Hamilton (Hamilton et al. 1987) in colorectal tissue slides. In that study, the authors used semi-automated image analysis methods to classify normal colorectal mucosa and adenocarcinoma, while in another paper automated image analysis was used (Hamilton et al. 1997). Hamilton et al. introduced an image texture analysis method to locate dysplastic fields in colorectal samples (Hamilton et al. 1997). The automatic identification of focal areas of colorectal dysplasia was based on cooccurrence matrix and optical density at low power microscopic images.

This study also showed that the combination of automated localization at low magnification and knowledge-based image segmentation at high magnification creates an automated tool for supporting diagnostic decision making. Using image texture analysis the authors achieved 83% accuracy in correctly classifying mucosa either as normal or dysplastic. The following studies attained an even higher accuracy of success, reaching 90%, or more (Esgiar et al. 1998b,a). Among the classification strategies put to the task, both linear discriminate function and k-nearest-neighbour non-parametric classifiers were separately used to identify cancerous colonic mucosa (Esgiar et al. 1998b).

In another study, Esgiar et. al. used gray level co-occurrence matrices, like many other studies, with non-overlapping square windows along an image and for each window, four co-occurrence matrices were calculated at four angles (0°, 45°, 90°, and 135°) from each image (Esgiar et al. 1998a).

Esgiar et. al. continued the study of classifying tissues samples taken from colons with colorectal cancer or diverticulosis calculating fractal dimensions (Esgiar et al. 2002). They discussed that fractal analysis did add a small improvement to the results obtained using correlation and entropy alone. They also reported that the reason for the relatively small improvement was that fractal dimension had been shown to be highly coupled with both correlation and entropy features. Hence, they concluded that, their research highlighted the need for researchers to find techniques which add independent value to other analysis techniques, such as, additional image sections, e.g. those transverse image sections perpendicular to those normally examined and the analysis of color.

Esgiar et. al. used a "leave-one-out" approach to obtain nearly unbiased estimates of classification error rates. This method removes one observation from $N$ observations and treats the remaining $N - 1$ as a training set. The one left out is then classified. This technique is then repeated for the $N$ training observations. Classification was achieved using the "nearest neighbour non-parametric" classification analysis with $k = 2$.

In another study, Filippas et. al. focused on the identification of normal and cancerous colonic mucosa using a genetic algorithm (Filippas et al. 2003). Filippas et. al. performed classification on three different family of features; histogram based features, features derived from grey-level difference statistics and co-occurrence matrices. This study had very promising accuracy results, 100% accuracy in the classification of the images in the training set and up to 91% in that of the test set. On the other hand, it should also be notified that these experiments were performed with 31 images in total (20 images in the training set and 11 images images in a test set) and increasing the number of images in the experimental dataset would probably change the resultant accuracy levels.

In a couple of following colon histopathology studies, hyperspectral imagery data was processed (Nasir et al. 2004, Masood et al. 2006). The microscopic level images of human colon tissue cells were acquired using hyperspectral imaging technology at contiguous wavelength intervals of visible light. Note that, while hyperspectral imagery data provides a wealth of information, its large size normally means high computational processing complexity.

Nasir et.al. and Masood et.al. used a support vector machine (SVM) classifier, to discriminate between normal and malignant colon tissue cells. Nasir et. al. performed segmentation for the parts of the colon tissue i.e. nuclei, cytoplasm, lamina propria, and lumen. They performed SVM classification on morphological and several statistical features separately. They acquired classification performances from 86% to 89% in various experiments. Similarly, Masood et.al. segmented the parts of the colon tissue into four parts; nuclei, cytoplasm, gland secretions and lamina propria. Furthermore, they calculated the morphological features that describe the shape, size, orientation and other geometrical attributes of the cellular components. Masood et.al. reached classification accuracy levels equal to or above 90%.

Nwoye et. al. used a fuzzy neural network classifier to perform differentiation between normal colon polyps and adenocarcinomas in colorectal tissue slides (Nwoye et al. 2006). They reached accuracy level up to 96.5% by using both spectral and gray scale statistical co-occurrence matrix analysis of the microscopic cell images. Fourier spectral extraction parameters measured in that study included spectral entropy, energy and inertia, while from co-occurrence matrix the statistical features, statistical contrast, entropy, moment and correlation were calculated. The authors reported that the novelty of the algorithm was the independence of the feature extraction procedure adopted which is also the one of the highlights of our proposed framework.

Ficsor et. al. experimented automated classification of normal, colorectal dysplasia, aspecific colitis and ulcerative colitis for routinely processed H&E stained high resolution (0.24 mm/pixel) histological sections (Ficsor et al. 2008). They performed detection of nuclei, tissue components, and structures yielding several cytometric morphological parameters. They experimented leave-one-out discriminant analysis method for classification of the sample groups. They found out that cellular morphometric features showed no significant differences in these benign colon alterations, however, gland related morphological differences for normal mucosa, ulcerative colitis, and aspecific colitis did. As a result, they reached to the overall classification accuracy of 88%.

Finally, a preliminary form of this thesis research being based on a limited number of colon histopathology slide image database was published (Önder et al. 2010).

## 3.3 Classification Accuracy Evaluation

In this section, the output classification accuracy levels obtained in several histopathological image analysis studies are listed. The resultant classification accuracy levels are evaluated separately as colon based studies and studies on other tissue types.

### 3.3.1 Accuracy Evaluation In Colon Tissues

The classification accuracy levels obtained in several studies performed on colonic tissues reached up to; the 83% of test images were correctly classified to locate dysplastic fields in colorectal samples (Hamilton et al. 1997), the overall classification accuracy of 88% in the classification of normal, colorectal dysplasia, aspecific colitis and ulcerative colitis was achieved (Ficsor et al. 2008) .

Being specific to the colorectal cancers, accuracy levels reached up to; the results were confirmed with the test set an overall accuracy of 90.2% (Esgiar et al. 1998b), up to 89% (Nasir et al. 2004), Masood et.al. reached classification accuracy levels equal to or above 90% (Masood et al. 2006), the 96.5% overall classification rate was achieved (Nwoye et al. 2006) In addition, (Filippas et al. 2003) reported very promising accuracy results, 100% accuracy in the classification of the images in the training set and up to 91% in that of the test set.

These accuracy levels of colon tissue researches could also be compared with our resultant labelling and classification performances presented in Section 5.7.5 and Section 5.10.

### 3.3.2 Accuracy Evaluation In Non-colonic Tissues

Features derived from segmented nuclei and glands from histopathology are usually a prerequisite to extract higher level information regarding the state of the disease. For instance, the grading of prostate cancer by Jafari-Khouzani and Soltanian-Zadeh yielded 97% accuracy for H&E stained imagery based on features derived from nuclear structures in (Jafari-Khouzani and Soltanian-Zadeh 2003).

The classification of histopathology imagery using spatial architecture information as presented in Weyn et al. resulted with 88.7% - 96.8% accuracy in the diagnosis of lung cancer, 94.9% accuracy in the typing of malignant mesothelioma, and 80.0% - 82.9% accuracy in the prognosis of malignant mesothelioma for Feulgen-stained lung sections (Weyn et al. 1998).

Analysis of Feulgen-stained breast tissue sections by Van de Wouwer et al. reached 67.1% accuracy in classifying nuclei as benign or malignant, but 100% classification on patient level (Wouwer et al. 2000). Tabesh et al. found 96.7% accuracy in discriminating between prostate tissue slides with cancer and no cancer, and 81% accuracy in the discrimination between low and high Gleason grades in the same imagery (Tabesh et al. 2007).

The analysis of H&E stained brain tissue by Demir et al. gave 95.5% - 97.1% accuracy in the discrimination between benign and cancerous tissue (Demir and Yener 2006). In another research, Keenan et al. reported accuracies of 62.3% -

76.5% in the grading of H&E stained cervical tissue (Keenan et al. 2000).

## 3.4 Review Of The Quasi-supervised Learning

Quasi-supervised Learning (QSL) is a statistical learning algorithm that contrasts two datasets by computing estimates for the posterior probabilities of their samples of belonging in either dataset (Karaçalı 2010). The QSL method addresses a target identification problem where labelled samples are available from one class only, in a control dataset. A second, unlabelled dataset is also provided and contains a mixture of samples from both control and target classes.

For biomedical data analysis tasks, it is quite easy to collect dataset of control samples easily while representative abnormalities require laborious manual identification. Therefore, the proposed strategy accommodates the biomedical data analysis task well.

In performance evaluation experiments on synthetic target detection data, QSL method outperformed alternative strategies based on SVM classification and minimum spanning trees for varying dataset size, overlap, and dimensionality (Karaçalı 2010).

QSL algorithm has been successfully used in the preliminary research of this dissertation, based on a limited number of colon histopathology slide image database (Önder et al. 2010). Köktürk also studied the separation of the electroencephalography data recorded under different visual stimuli by using the quasi-supervised learning algorithm (Köktürk 2011). The data used in that study contained multiple channel EEG recording under six different visual stimuli in random successive order. Köktürk identified condition-specific EEG profiles in different comparison scenarios by using standard binary QSL and also extending its M-ary version. The results revealed that QSL algorithm was efficient in capturing the distinction between the experimental data samples.

QSL could also have application in many other areas rather than biomedical data analysis. Güven has performed a study on aerial images using the contrasted information between natural and man-made structures (Güven 2010). The main purpose of that study was the detection of man-made structures or differences on the terrain as a result of habitation. Güven has calculated Haralick texture features based on gray level co-occurrence matrices. The results showed that QSL algorithm was able to identify the indicators of human presence in a region such as houses, roads and objects that are not likely to be observed in areas free from human habitation.

# CHAPTER 4

# METHODS

In this chapter, the background information for the elementary parts of the proposed framework is presented. The first section introduces the feature vectors and details the feature extraction geometry. In the second section, the core algorithm of our proposed framework, QSL, is presented with basic derivations in a sequenced manner. The third section devoted to the SVM supervised classifier along with a concrete implementation of SVM, $SVM^{light}$. In the later section, several alternative methods of vector dimensionality reduction are described. Following the major feature selection methodologies, several dimensionality reduction methods based on feature extraction, namely, Principal Component Analysis, Isomap, Locally Linear Embedding and Independent Component Analysis including $FastICA$ algorithm are detailed. In the last section, the basic overview of the proposed methodology is presented as an integrated labelling system.

## 4.1 Texture Feature Vector Extraction

One of the classical approaches to texture classification is to use texture features derived from co-occurrence matrices (Haralick et al. 1973). A co-occurrence matrix is a local approximation to the joint distribution of gray level values of pixel pairs at specified distances and orientations. In our study, the entry of a co-occurrence matrix $M_{I,d}(i,j)$ at the $i$'th row and the $j$'th column is calculated by;

$$M_{I,d}(i,j) = \frac{\sum_{p,q \epsilon B_r(x)} 1\{I(p)=i, I(q)=j, \rho(p,q)=d\}}{\sum_{p,q \epsilon B_r(x)} 1\{\rho(p,q)=d\}} \tag{4.1}$$

where $I$ is the gray level image in consideration, $d$ is the pairwise distance of pixels and $B_r(x)$ defines the neighbourhood of radius $r$ centered at an image pixel $x$. Furthermore, $p$ and $q$ represent two pixels in the image and $\rho(p,q)$ represents the Euclidean distance between them. In the definition above, no direction specific relation was enforced between $p$ and $q$ though the formalism allows for incorporating such considerations into the representation. The function $1\{\cdot\}$ is a binary function that takes the values 1 or 0 when its argument is true and false, respectively.

Although it is theoretically possible to calculate texture feature vectors around every pixel of an image under consideration, it is not computationally feasible. Therefore, we have assumed that the co-occurrence matrix varied smoothly across the image, and carried out feature vector computation for points on a regular grid with a spacing of $r$. Furthermore, the labelling decision made for a feature vector of a grid point was generalized to the square image region of size $r \times r$ centered at that point. The sketch for the calculation geometry is presented in Figure 4.1.

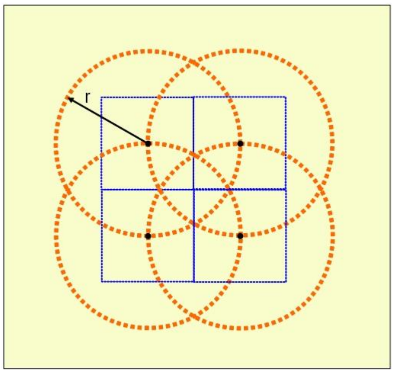


Figure 4.1. Texture feature vector calculation geometry.

Texture features that characterize the appearance of an image square are composed of the first order and the second order characteristics. The first order texture features were generated using local histograms, whereas the second order features were generated from co-occurrence matrices calculated in a local image region. The list of texture features used in this study is given below. The second order texture feature characteristics except (2.n) are known as Haralick features (Haralick et al. 1973).

1. First Order Features (Pratt 1991, Gonzalez and Woods 1992)

   (a) Mean Value of Pixels,

   (b) Variance,

   (c) Skewness,

   (d) Kurtosis,

   (e) Entropy,

   (f) Energy,

   (g) Maximum Value of Histogram,

   (h) Corresponding Pixel Value for Maximum Value of Histogram (CPVMVH).

2. Second Order Features

   (a) Angular Second Moment,

   (b) Contrast,

   (c) Correlation,

   (d) Sum of Squares: Variance,

   (e) Inverse Difference Moment,

   (f) Sum Average,

   (g) Sum Entropy,

   (h) Sum Variance,

   (i) Entropy,

   (j) Difference Variance,

   (k) Difference Entropy,

   (l) Information Measures of Correlation 1 (IMC1),

   (m) Information Measures of Correlation 2 (IMC2),

   (n) Maximum Probability (maximum co-occurrence matrix element).

For a local image square, each texture feature in the list above constitutes a component of the corresponding texture feature vector. It should be noted that for a local image square, various second order texture features can be calculated for co-occurrence matrices generated using different pairwise pixel distance $d$ values. The second order feature vector components calculated using different $d$ values can then be combined with the first order features to form a more detailed feature vector.

Another strategy to enrich the set of texture features is to use a hierarchical organization. In a hierarchical computation of texture features, a set of radius values that are multiples of $r$ is used to define a nested set of neighbourhoods. Texture feature vector extraction is performed separately for each neighbourhood and these vectors are then concatenated in order to have a higher dimensional texture feature vector. In this strategy, the $(i, j)$'th entry of a co-occurrence matrix of a hierarchical level $h = 1, 2, .., H$ is calculated by:

$$M_{I,h,d}(i,j) = \frac{\sum_{p,q \epsilon B_{r\times h}(x)} 1\{I(p) = i, I(q) = j, \rho(p,q) = d\}}{\sum_{p,q \epsilon B_{r\times h}(x)} 1\{\rho(p,q) = d\}}, \tag{4.2}$$

similar to Equation 4.1. The difference is the size of neighbourhood $B_{r\times h}(x)$. Note that $h = 1$ corresponds to non-hierarchical texture feature extraction. The idea behind hierarchical texture feature vectors is to identify small scale texture characteristics together with those present at larger scales. The geometry of hierarchical texture feature calculation for $h = \{1, 2\}$ is shown in Figure 4.2.

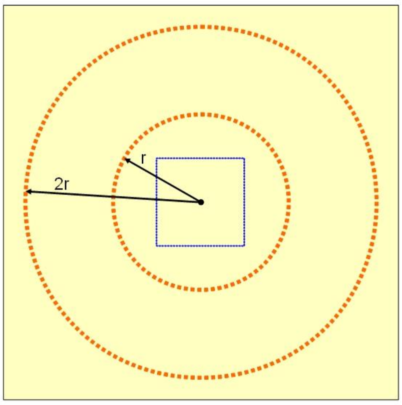


Figure 4.2. Hierarchical texture feature vector calculation geometry for $h = \{1, 2\}$.

## 4.2 Quasi-supervised Learning

Quasi-supervised Learning (QSL) is a statistical learning algorithm that contrasts two datasets by computing estimates for the posterior probabilities of their samples of belonging in either dataset (Karaçalı 2010). The QSL method addresses for a target identification problem where labelled samples are available from one class only, in a control dataset. A second, unlabelled dataset is also provided and contains a mixture of samples from both control and target classes. Identification is carried out using a computational algorithm that contrasts the unlabelled mixed dataset samples to the control dataset, and selecting those that are dissimilar from the control samples beyond a statistical significance level as target samples.

The lack of class labels other than the control dataset changes this problem more towards unsupervised learning rather than semi-supervised learning (Chapelle et al. 2006). For biomedical data analysis tasks, it is typically very easy to collect a dataset of control samples while representative abnormalities require laborious manual identification. Therefore, the proposed strategy accommodates the biomedical data analysis task well. In computational analysis of histology slides, for instance, a pathologist can easily identify tissue cross-sections that are free from cancerous abnormalities. However, such abnormalities may occur amid tissue that is benign in appearance and have to be painstakingly labelled by an expert pathologist.

In a wider perspective, estimation of posterior probabilities from available data makes QSL algorithm suitable for classification as a model-free alternative to existing techniques. Especially, in cases where the data do not allow perfect separation of the different classes, the algorithm can be expected to outperform off-the-shelf classification algorithms as it will not search for a separation boundary optimized according to some criteria. In performance evaluation experiments on synthetic target detection data, the QSL method outperformed alternative strategies based on Support Vector Machine classification and minimum spanning trees for varying dataset size, overlap, and dimensionality (Karaçalı 2010).

### 4.2.1 Likelihood Ratio Estimation Via The Nearest Neighbour Rule

Given a reference set $R = \{\mathbf{x_i}, y_i\}$ of points $\mathbf{x_i} \in X$ and their respective class labels $y_i \in \{0, 1\}$ for $i = 1, 2, \ldots, \ell$, a nearest neighbour classifier is defined by

$$F_R(\mathbf{x}) = y_{i_0} \text{ where } i_0 = \arg_{i=1,2,\ldots,\ell} \min d(\mathbf{x}, \mathbf{x_i}) \tag{4.3}$$

for $\mathbf{x} \in X$, and where $d(.,.)$ denotes the metric on $X$.

The nearest neighbour classifier in Equation 4.3 has been a benchmark classification method in the pattern recognition literature. It is quite simple and has asymptotic properties linking its error rate to that of the optimal Bayes rate (Duda et al. 2000). Indeed, it can be shown that the asymptotic error rate of the nearest neighbour classifier is bounded from above by twice the Bayes rate.

Now, the ratio of the classification decisions, for a point $\mathbf{x}$ will be considered during the course of successive classifications each time using a different reference set, as the number of classifications grows large.

Let $\{R_j\}$, $j = 1, 2, \ldots, M$, be a collection of independent and identically distributed reference sets, consisting of $n$ points from each of the two classes. Define $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ by

$$f_0(\mathbf{x}) = \frac{\sum_{j=1}^{M} 1(F_{R_j}(\mathbf{x}) = 0)}{M} \tag{4.4}$$

and

$$f_1(\mathbf{x}) = \frac{\sum_{j=1}^{M} 1(F_{R_j}(\mathbf{x}) = 1)}{M} \tag{4.5}$$

where 1(statement) is 1 if statement holds and 0 otherwise. The critical observation is that for sufficiently large $M$,

$$f_0(\mathbf{x}) \simeq \frac{p(\mathbf{x}|\mathbf{x} \in C_0)}{p(\mathbf{x}|\mathbf{x} \in C_0) + p(\mathbf{x}|\mathbf{x} \in C_1)} \tag{4.6}$$

and

$$f_1(\mathbf{x}) \simeq \frac{p(\mathbf{x}|\mathbf{x} \in C_1)}{p(\mathbf{x}|\mathbf{x} \in C_0) + p(\mathbf{x}|\mathbf{x} \in C_1)} \tag{4.7}$$

providing

$$\frac{f_0(\mathbf{x})}{f_1(\mathbf{x})} \simeq \frac{p(\mathbf{x}|\mathbf{x} \in C_0)}{p(\mathbf{x}|\mathbf{x} \in C_1)} \tag{4.8}$$

where $p(\mathbf{x}|\mathbf{x} \in C_0)$ and $p(\mathbf{x}|\mathbf{x} \in C_1)$ represent the class conditional probability densities for the classes $C_0$ and $C_1$, respectively (Fukunaga and Hostetler 1975, Duda et al. 2000).

Note also that since the inclusion of an equal number of samples from each class in the reference set leads to equal prior probabilities for $C_0$ and $C_1$, in effect, $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ in Equations 4.6 and 4.7 compute estimates of the posterior distributions of the classes $C_0$ and $C_1$ at the point $\mathbf{x}$.

This observation suggests that given a point $\mathbf{x} \in X$, the likelihood ratio as well as the posterior probabilities of the two classes at $\mathbf{x}$ can be estimated based on a dataset $\{\mathbf{x_i}, y_i\}$ by carrying out multiple nearest neighbour classifications on $\mathbf{x}$ using randomly chosen reference sets from $\{\mathbf{x_i}\}$ with equal representation from both classes, and keeping track of the number of times $\mathbf{x}$ is assigned to $C_0$ and to $C_1$.

The accuracy of the above estimate of the log-likelihood ratio at a point $\mathbf{x}$ clearly depends on $M$, the number of successive random nearest neighbour classifications. With $n$ points from each class in the reference set, the total number of distinct reference sets of size $2n$ is given by

$$\binom{\ell_0}{n}\binom{\ell_1}{n} \tag{4.9}$$

where $\ell_0$ and $\ell_1$ denote the number of points in the set $\{\mathbf{x_i}\}$ belonging to the respective classes. Even for modest reference set sizes, say of 100 samples each, the number of distinct nearest neighbour classifications that should be carried out reaches levels over $10^{50}$, well beyond the performance of the present day computing equipment.

While carrying out an exhaustive evaluation of all possible random nearest neighbour classifications is not feasible, it is still possible to compute the average number of times a given point would be assigned to either class at the end of such an evaluation.

Consider the distances $d_i = d(\mathbf{x}, \mathbf{x_i})$ between a given point $\mathbf{x}$ and each $\mathbf{x_i}$ for $i = 1, 2, \ldots, \ell$. Let $d_{(i)}$ denotes the ordered sequence of all $\{d_i\}$ with $d_{(1)} \leq d_{(2)} \ldots \leq d_{(\ell)}$ and $\{\mathbf{x_{(i)}}\}$ and $\{y_{(i)}\}$ be such that $d_{(i)} = d(\mathbf{x}, \mathbf{x_{(i)}})$ and $y_{(i)}$ is the class label of

$\mathbf{x_{(i)}}$. After an exhaustive nearest neighbour analysis, $f_0(\mathbf{x})$ represents the probability $Pr(y = 0)$ of assigning $\mathbf{x}$ to the class $C_0$ based on a reference set $R$ with $n$ points from both classes selected randomly from the $\{\mathbf{x_i}\}$,

$$f_0(\mathbf{x}) = Pr(y = 0) \tag{4.10}$$

Note that the term $Pr(y = 0)$ can be decomposed conditionally on whether or not the point $\mathbf{x_{(1)}}$ is in $R$ in as

$$f_0(\mathbf{x}) = Pr\{\mathbf{x_{(1)}} \in R\}1(y_{(1)} = 0) + Pr\{\mathbf{x_{(1)}} \notin R\}Pr\{y = 0|\mathbf{x_{(1)}} \notin R\} \tag{4.11}$$

since $Pr\{y = 0|\mathbf{x_{(1)}} \in R\}$ is 1 if $y_{(1)} = 0$ and 0 otherwise. For notational simplicity, we can define $E_k$ as the joint event $\mathbf{x_{(1)}}, \mathbf{x_{(2)}}, \ldots, \mathbf{x_{(k)}} \notin R$. Carrying the same decomposition strategy further to $Pr\{y = 0|E_1\}$ provides;

$$\begin{aligned} Pr\{y = 0|E_1\} = Pr\{\mathbf{x_{(2)}} \in R|E_1\}1(y_{(2)} = 0) + \\ Pr\{\mathbf{x_{(2)}} \notin R|E_1\}Pr\{y = 0|E_2\} \end{aligned} \tag{4.12}$$

and in general:

$$\begin{aligned} Pr\{y = 0|E_{k-1}\} = Pr\{\mathbf{x_{(k)}} \in R|E_{k-1}\}1(y_{(k)} = 0) + \\ Pr\{\mathbf{x_{(k)}} \notin R|E_{k-1}\}Pr\{y = 0|E_k\} \end{aligned} \tag{4.13}$$

Furthermore $R$ must have at least $2n$ data points the decomposition does not need to be carried out beyond some $k^*$ given by

$$k^* = max\{k|\sum_{k'=k}^{\ell} 1(y_{(k')} = 0) \geq n \; and \; \sum_{k'=k}^{\ell} 1(y_{(k')} = 1) \geq n\} \tag{4.14}$$

since $Pr\{\mathbf{x_{(k^*)}} \in R|E_{k^*-1}\} = 1$ and $Pr\{\mathbf{x_{(k^*)}} \notin R|E_{k^*-1}\} = 0$. The algebraic development above can be repeated for $f_1(\mathbf{x})$ in an identical manner.

The following algorithm elucidates the computation all steps of $f_L(\mathbf{x})$ when $L \in \{0, 1\}$ and $\mathbf{x}$ is based on $\{\mathbf{x_i}, y_i\}$ where $i = 1, 2, \ldots, \ell$ and a fixed $n$.

1. Calculate $d_i = d(\mathbf{x}, \mathbf{x_i})$.

2. Sort $\{d_i\}, d_{(1)} \leq d_{(2)} \ldots \leq d_{(\ell)}$ and determine the $\{\mathbf{x_i}\}$ and $\{y_i\}$ according to sorted distances.

3. Determine $k^*$ in 4.14 and set $Pr\{y = L|E_{k^*-1}\} = 1(y_{(k^*)} = L)$.

4. Calculate $Pr\{y = L|E_k\} = Pr\{\mathbf{x_{(k)}} \in R|E_k\}1(y_{(k+1)} = L) + Pr\{\mathbf{x_{(k)}} \notin R|E_k\}Pr\{y = L|E_{k+1}\}$ for $k = k^* - 1, k^* - 2, \ldots, 1$.

5. Calculate $f_L(\mathbf{x}) = Pr\{\mathbf{x_{(1)}} \in R\}1(y_{(1)} =) + Pr\{\mathbf{x_{(1)}} \notin R\}Pr\{y = L|E_1\}$

Note that $Pr\{\mathbf{x_{(k)}} \in R|E_{k-1}\}$ can be calculated by

$$Pr\{\mathbf{x_{(k)}} \in R|E_{k-1}\} = 1 - Pr\{\mathbf{x_{(k)}} \notin R|E_{k-1}\} = 1 - \frac{\binom{\ell_0^{k+1}}{n}\binom{\ell_1^{k+1}}{n}}{\binom{\ell_0^{k}}{n}\binom{\ell_1^{k}}{n}} \tag{4.15}$$

where $\ell_0^k$ represents the number of points that belong to $C_0$ in the set $\{\mathbf{x_{(k)}}, \mathbf{x_{(k+1)}}, \ldots, \mathbf{x}_{(\ell)}\}$ and denotes the number of points in the same set that belong to $C_1$ with,

$$\ell_0^k = \sum_{i=k}^{\ell} 1(y_{(i)} = 0),\ \ell_1^k = \sum_{i=k}^{\ell} 1(y_{(i)} = 1) \tag{4.16}$$

Since for $y_{(k)} = 0$ we have $\ell_0^{k+1} = \ell_0^k - 1$ and $\ell_1^{k+1} = \ell_1^k$, and similarly for $y_{(k)} = 1\ell_0^{k+1} = \ell_0^k$ and $\ell_1^{k+1} = \ell_1^k - 1$, we obtain:

$$Pr\{\mathbf{x_{(k)}} \in R|E_{k-1}\} = \begin{cases} \frac{n}{\ell_0^k} & \text{if } \ y_{(k)} = 0 \\ \frac{n}{\ell_1^k} & \text{if } \ y_{(k)} = 1 \end{cases} \tag{4.17}$$

### 4.2.2 Class Overlap Measures

In classification problems, class-overlap measures provide information about the separability of the classes. A successful classification, in particular, can be constructed with smaller class overlap measures.

There are several measures of class overlap that can be computed using the estimated posterior probabilities. One of them is $M_{LLR}(\mathbf{x})$ that computes the log-likelihood ratio of two classes at a point $\mathbf{x} \in \chi$, given by

$$M_{LLR}(\mathbf{x}) = \log(\frac{f_0(\mathbf{x})}{f_1(\mathbf{x})}) \tag{4.18}$$

using the expressions for $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ given in Eqn. 4.6 and 4.7 for all $\mathbf{x}$ with $f_0(\mathbf{x}) \neq 0$ and $f_1(\mathbf{x}) \neq 0$. When the two classes are mostly overlap the ratio $f_0(\mathbf{x})/f_1(\mathbf{x})$ goes to 1 and $M_{LLR}(\mathbf{x})$ goes to 0. The major benefit of the $M_{LLR}(\mathbf{x})$ is it gives opportunity to determine which class $\mathbf{x_i}$ belongs to among two classes $C_0$ or $C_1$.

Another way of measuring the class-overlap or similarity between distributions is using the Henze-Penroze affinity measure, the measure that computes the integral,

$$\int_x \frac{2p_1(\mathbf{x})p_2(\mathbf{x})}{p_1(\mathbf{x}) + p_2(\mathbf{x})} dx \tag{4.19}$$

for any given probability distributions $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ (Henze and Penrose 1999, Neemuchwala and Hero 2005). The integral goes to 1 when $p_1(\mathbf{x}) = p_2(\mathbf{x})$ for all $\mathbf{x}$. We define the measure $M_{HP-like}(\mathbf{x})$ for a sample $\mathbf{x}$ as a variant of the integrand above by

$$M_{HP-like}(\mathbf{x}) = f_0(\mathbf{x})f_1(\mathbf{x}) \simeq \frac{p(\mathbf{x}|\mathbf{x} \in C_0)p(\mathbf{x}|\mathbf{x} \in C_1)}{(p(\mathbf{x}|\mathbf{x} \in C_0) + p(\mathbf{x}|\mathbf{x} \in C_1))^2} \tag{4.20}$$

Note that over the regions of overlap, $f_0(\mathbf{x}) \simeq f_1(\mathbf{x}) \simeq 1/2$ and $M_{HP-like}(\mathbf{x})$ approaches 1/4. Conversely, for points that are highly specific to one or the other class, $M_{HP-like}(\mathbf{x})$ is near zero.

A final measure of overlap can be computed using the difference of $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ by

$$M_{Diff}(\mathbf{x}) = f_0(\mathbf{x}) - f_1(\mathbf{x}) \simeq \frac{p(\mathbf{x}|\mathbf{x} \in C_0) - p(\mathbf{x}|\mathbf{x} \in C_1)}{p(\mathbf{x}|\mathbf{x} \in C_0) + p(\mathbf{x}|\mathbf{x} \in C_1)} \tag{4.21}$$

This measure is similar to $M_{LLR}(\mathbf{x})$ in the sense that the points of strong overlap are also given by the set of points for which $M_{Diff} \simeq 0$. On the other hand, $M_{Diff}$ can be computed for any $\mathbf{x}$, even those for which $f_0(\mathbf{x}) = 0$ or $f_1(\mathbf{x}) = 0$.

### 4.2.3 Selection Of The Optimal Reference Set Size

The accuracy of the algorithm that estimates posterior distributions is related to reference set size, because the samples in the reference set have a role in the determination of the posterior distributions. Recall that $n$ denotes the number of samples from each class in the reference set. Ideally; the best $n$ provides the minimum class overlap, equivalently, maximal class separation. In addition, $n$ must be chosen as small as possible to avoid too flexible nearest neighbour classification and to reduce estimation noise (Karaçalı and Krim 2003, Karaçalı et al. 2004). In light of these considerations, the following cost functional is proposed;

$$E(n) = 4\sum_{i}(p_0(\mathbf{x})p_1(\mathbf{x})) + 2n \tag{4.22}$$

to be minimized with respect to $n$. The first term in this cost functional represents the penalty for the large class overlaps, while, the second term represents the preference for smaller $n$ to achieve better generalisation with nearest neighbourhood classification. The scaling of the first term by a factor of 4 ensures that the costs incurred in the two marginal scenarios where, at the one end, $M_{HP-like} = 1/4$ for all $\mathbf{x_i}$ indicating complete overlap, and at the other $n = \ell/2$ when the reference sets are as large as they can be (assuming $\ell_0 = \ell_1$), are equal. The experimental verification of the cost functional was given in great detail in (Karaçalı 2010).

## 4.3 Support Vector Machines

The Support Vector Machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis (Cortes and Vapnik 1995, Vapnik 1998, Burges 1998). The standard SVM, a non-probabilistic binary linear classifier, takes a set of input data and assigns the sample in question to one of the two possible categories. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

More formally, SVM constructs a hyperplane in a high or infinite dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

The next section details the problem of classification for linearly separable data and introduces the concept of margin and the essence of SVM margin maximization. The methodology of SVM is then extended to data which is not fully linearly separable and to the non-linear case.

### 4.3.1 Linearly Separable Binary Classification

Suppose we have $L$ training points, where each input $\mathbf{x_i}$ has $D$ attributes (i.e. is of dimensionality $D$) and is in one of two classes $y_i$ = -1 or +1, i.e our training data is of the form: $\{\mathbf{x_i}, y_i\}$ where $i = 1, \dots, L$, $y_i\{-1, 1\}$, $\mathbf{x} \in R^D$

It is assumed that the data is linearly separable, meaning that a line can be drawn on a graph of $\mathbf{x_1}$ versus $\mathbf{x_2}$ separating the two classes when $D = 2$ and a hyperplane on graphs of $x_1, x_2, \dots x_D$ for when $D > 2$. This hyperplane can be described by $\mathbf{w} \cdot \mathbf{x} + b = 0$ where;

1. $\mathbf{w}$ is normal to the hyperplane.

2. $\cdot$ operator represents the inner product of two vectors.

3. $\frac{b}{\|\mathbf{w}\|}$ is the perpendicular distance from the hyperplane to the origin.

Support Vectors are the examples closest to the separating hyperplane and the aim of Support Vector Machines (SVM) is to orientate this hyperplane in such a way as to be as far as possible from the closest members of both classes, as represented in Figure 4.3. In this diagram, circles represent the Support Vectors.

Referring to the figure, implementing an SVM boils down to selecting the

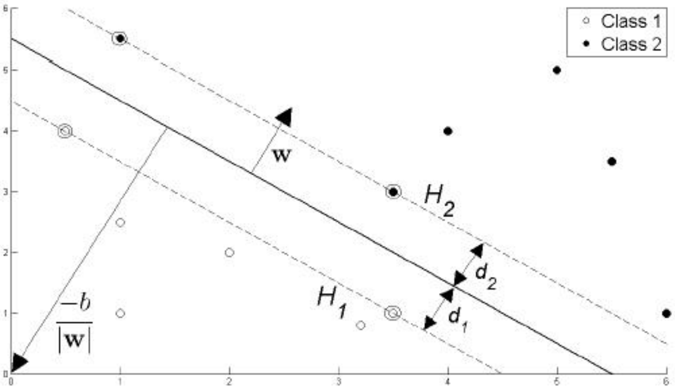


Figure 4.3. Hyperplane through two linearly separable classes.

variables **w** and b so that our training data can be described by;

$$\mathbf{x_i} \cdot \mathbf{w} + b \geq +1 \, for \, y_i = +1 \tag{4.23}$$

$$\mathbf{x_i} \cdot \mathbf{w} + b \leq -1 \, for \, y_i = -1 \tag{4.24}$$

which also can be combined into:

$$y_i(\mathbf{x_i} \cdot \mathbf{w} + b) - 1 \geq 0 \, \forall_i \tag{4.25}$$

Considering the points that lie closest to the separating hyperplane, i.e. the Support Vectors (shown in circles in the diagram), then the two planes $H_1$ and $H_2$ that these points lie on can be described by:

$$\mathbf{x_i} \cdot \mathbf{w} + b = +1 \, for \, H_1 \tag{4.26}$$

$$\mathbf{x_i} \cdot \mathbf{w} + b = -1 \, for \, H_2 \tag{4.27}$$

Referring to Figure 4.3, we define $d_1$ as being the distance from $H_1$ to the hyperplane and $d_2$ from $H_2$ to it. The hyperplane's equidistance from $H_1$ and $H_2$ means that $d_1 = d_2$ that is a quantity known as the SVM's margin. In order to orient the hyperplane to be as far from the Support Vectors as possible, this margin is to be maximised.

Simple vector geometry shows that the margin is equal to $\frac{1}{\|\mathbf{w}\|}$ and maximizing it subject to the constraint in Equation 4.25 is equivalent to finding: minimum $\|\mathbf{w}\|$ such that $\mathbf{y_i} \cdot \mathbf{w} + b - 1 \geq 0 \, \forall_i$.

Minimizing $\|\mathbf{w}\|$ is equivalent to minimizing $\frac{1}{2}\|\mathbf{w}\|^2$ and the use of this term makes it possible to perform Quadratic Programming (QP) optimization later on. Therefore it should be found:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \text{ s.t. } y_i(\mathbf{x_i} \cdot \mathbf{w} + b) - 1 \geq 0 \quad \forall_i \tag{4.28}$$

Now, the Lagrangian formulation of the problem has to be considered. There are two reasons for this; the first is that the constraint in Equation 4.25 will be replaced by constraints on the Lagrange multipliers themselves, which will be much easier to handle. The second is that in this reformulation of the problem, the training data will only appear (in the actual training and test algorithms) in the form of dot products between vectors. This is a crucial property which will allow us to generalize the procedure to the non-linear case (Section 4.3.3).

Thus, it is introduced that the positive Lagrange multipliers $\alpha_i$, $i = 1, \ldots, L$, one for each of the inequality constraints 4.25. Recall that the rule is that for constraints of the form $c_i \geq 0$, the constraint equations are multiplied by positive Lagrange multipliers and subtracted from the objective function, to form the Lagrangian. For equality constraints, the Lagrange multipliers are unconstrained. This gives Lagrangian:

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{L} \alpha_i(y_i(\mathbf{x_i} \cdot \mathbf{w} + b) - 1) \tag{4.29}$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{L} \alpha_i y_i(\mathbf{x_i} \cdot \mathbf{w} + b) + \sum_{i}^{L} \alpha_i \tag{4.30}$$

with the Lagrange multipliers $\alpha$, where $\alpha_i \geq 0 \, \forall_i$ .

We wish to find the $\mathbf{w}$ and b which minimizes, and the $\alpha$ which maximizes Equation 4.30 (whilst keeping $\alpha_i \geq 0 \, \forall_i$). This can be performed by differentiating

$L_P$ with respect to $\mathbf{w}$ and b and setting the derivatives to zero:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{L} \alpha_i y_i \mathbf{x_i} \tag{4.31}$$

$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{4.32}$$

Substituting Equation 4.31 and 4.32 into 4.30 gives a new formulation which, being dependent on $\alpha$, we need to maximize:

$$L_D = \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} \tag{4.33}$$

$$= \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i H_{ij} \alpha_j \; where \; H_{ij} = y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} \tag{4.34}$$

$$= \sum_{i=1}^{L} -\frac{1}{2} \alpha^T \mathbf{H} \alpha \; s.t. \; \alpha_i \geq 0 \, \forall_i, \; \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{4.35}$$

This new formulation $L_D$ is referred to as the dual form of the primary $L_P$. It is worth noting that the dual form requires only the dot product of each input vector $\mathbf{x_i}$ to be calculated, this is important for the Kernel Trick described in Section 4.3.3. Having moved from minimizing $L_P$ to maximizing $L_D$ is a convex quadratic optimization problem, and a QP solver is run which will return $\alpha$ and Equation 4.31 will give us $\mathbf{w}$. What remains is to calculate $b$.

Any data point satisfying Equation 4.32 which is a Support Vector $\mathbf{x_s}$ will have the form: $y_s(\mathbf{x_s} \cdot \mathbf{w}) + b$. Substituting in Equation 4.31 yields;

$$y_s(\sum_{m \in S} \alpha_m y_m \mathbf{x_m} \cdot \mathbf{x_s} + b) = 1 \tag{4.36}$$

where $S$ denotes the set of indices of the Support Vectors. $S$ is determined by finding the indices $i$ where $\alpha_i > 0$. Multiplying through by $y_s$ and then using $y_s^2 = 1$ from

Equations 4.23 and 4.24:

$$y_s^2(\sum_{m \in S} \alpha_m y_m \mathbf{x_m} \cdot \mathbf{x_s} + b) = y_s \tag{4.37}$$

$$b = y_s - \sum_{m \in S} \alpha_m y_m \mathbf{x_m} \cdot \mathbf{x_s} \tag{4.38}$$

Instead of using an arbitrary Support Vector $\mathbf{x_s}$, it is better to take an average over all of the Support Vectors in $S$;

$$b = \frac{1}{N_S} \sum_{m \in S} (y_s - \alpha_m y_m \mathbf{x_m} \cdot \mathbf{x_s}) \tag{4.39}$$

where the $N_S$ is the number of Support Vectors. We now have the variables $\mathbf{w}$ and $b$ that define our separating hyperplane's optimal orientation and hence the SVM.

Summarising the steps to be performed in order to use an SVM to solve a linearly separable, binary classification problem:

- Create $\mathbf{H}$, where $H_{ij} = y_i y_j \mathbf{x_i} \cdot \mathbf{x_j}$.
- Find $\alpha$ so that $\sum_{i=1}^{L} \alpha_i - \frac{1}{2}\alpha^T \mathbf{H} \alpha$ is maximized, subject to the constraints $\alpha_i \geq 0 \ \forall_i$ and $\sum_{i=1}^{L} \alpha_i y_i = 0$. This is performed using a QP solver.
- Calculate $\mathbf{w} = \sum_{i=1}^{L} \alpha_i y_i \mathbf{x_i}$.
- Determine the set of Support Vectors $S$ by finding the indices such that $\alpha_i > 0$.
- Calculate $b = \frac{1}{N_S} \sum_{m \in S} (y_s - \alpha_m y_m \mathbf{x_m} \cdot \mathbf{x_s})$.
- Classify each new point $\mathbf{x}'$ by evaluating $y' = sgn(\mathbf{w} \cdot \mathbf{x}' + b)$.

### 4.3.2 Binary Classification For Non-separable Data

In order to extend the SVM methodology to handle data that is not fully linearly separable (soft margin SVM, see Figure 4.4), the constraints for Equations 4.23 and 4.24 are relaxed slightly to allow for misclassified points. This is done by introducing a positive slack variable $\xi_i,\ i = 1, \ldots, L$;

$$\mathbf{x_i} \cdot \mathbf{w} + b \geq +1 \quad for\, y_i = +1 - \xi_i \tag{4.40}$$

$$\mathbf{x_i} \cdot \mathbf{w} + b \leq -1 \quad for\, y_i = -1 + \xi_i \tag{4.41}$$

which can be combined into:

$$y_i(\mathbf{x_i} \cdot \mathbf{w} + b) - 1 + \xi_i \geq 0 \quad where \quad \xi_i \geq 0 \quad \forall_i \tag{4.42}$$

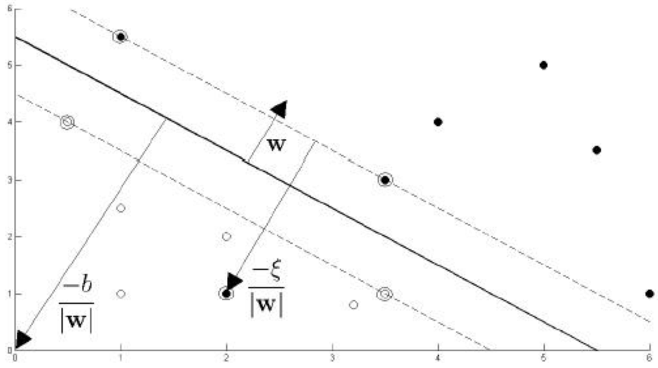


Figure 4.4. Hyperplane through two non-linearly separable classes.

In this soft margin SVM, data points on the incorrect side of the margin boundary have a penalty that increases with the distance from it. As we are trying to reduce the number of misclassifications, a sensible way to adapt our objective function Equation 4.28 is to find;

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{L}\xi_i \quad \text{s.t.} \quad \mathbf{y_i} \cdot \mathbf{w} + b - 1 + \xi_i \geq 0 \quad \forall_i \tag{4.43}$$

where the parameter $C$ controls the trade-off between the slack variable penalty and the size of the margin. Reformulating as a Lagrangian gives;

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{L}\xi_i - \sum_{i=1}^{L}\alpha_i(y_i(\mathbf{x_i} \cdot \mathbf{w} + b) - 1 + \xi_i) - \sum_{i=1}^{L}\mu_i\xi_i \tag{4.44}$$

that should be minimized with respect to $\mathbf{w}$, $b$ and $\xi_i$ and maximized with respect to $\alpha$ (where $\alpha_i \geq 0, \mu_i \geq 0$). Differentiating with respect to $\mathbf{w}$, $b$ and $\xi_i$ and setting the derivatives to zero:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{L} \alpha_i y_i \mathbf{x_i} \tag{4.45}$$

$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{4.46}$$

$$\frac{\partial L_P}{\partial \xi_i} = 0 \Rightarrow C = \alpha_i + \mu_i \tag{4.47}$$

Substituting these in, $L_D$ has the same form as Equation 4.35 before. However Equation 4.47 together with $\mu_i \geq 0 \quad \forall_i$, implies that $\alpha \leq C$. Therefore;

$$\max_{\alpha}(\sum_{i=1}^{L} \alpha_i - \frac{1}{2}\alpha^T \mathbf{H}\alpha) \quad s.t. \quad 0 \leq \alpha_i \leq C \quad \forall_i \quad and \quad \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{4.48}$$

should be found. $b$ is then calculated in the same way as in Equation 4.28 before, though in this instance the set of Support Vectors used to calculate $b$ is determined by finding the indices $i$ where $0 \leq \alpha_i \leq C$.

Summarising the steps to be performed in order solve a binary classification problem for data that is not fully linearly separable:

- Create $\mathbf{H}$, where $H_{ij} = y_i y_j \mathbf{x_i} \cdot \mathbf{x_j}$
- Find $\alpha$ so that $\sum_{i=1}^{L} \alpha_i - \frac{1}{2}\alpha^T \mathbf{H}\alpha$ is maximized, subject to the constraints $0 \leq \alpha_i \leq C \; \forall_i$ and $\sum_{i=1}^{L} \alpha_i y_i = 0$. (This is performed by using a QP solver.)
- Calculate $\mathbf{w} = \sum_{i=1}^{L} \alpha_i y_i \mathbf{x_i}$
- Determine the set of Support Vectors $S$ by finding the indices such that $0 \leq \alpha_i \leq C$.
- Calculate $b = \frac{1}{N_S} \sum_{m \in S}(y_s - \alpha_m y_m \mathbf{x_m} \cdot \mathbf{x_s})$.
- Classify each new point $\mathbf{x}'$ by evaluating $y' = sgn(\mathbf{w} \cdot \mathbf{x}' + b)$.

### 4.3.3 Non-linear Support Vector Machines

This section describes the generalisation of the above methods to the case where the decision function is not a linear function of the data. It was stated that a rather old trick, which is known as the Kernel Trick (Aizerman et al. 1964), can be used to accomplish this in a straightforward way (Boser et al. 1992). Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher dimensional space, presumably making the separation easier in that space. First notice that the only way in which the data appears in the training problem, Equation 4.48, is in the form of dot products, $\mathbf{x_i} \cdot \mathbf{x_j}$ . Now suppose the data is first mapped to some other (possibly infinite dimensional) Euclidean space, using a mapping which we will call $\Phi : \mathbf{x} \mapsto \Phi(\mathbf{x})$

Then, the training algorithm would only depend on the data through dot products in $\Phi(x)$, i.e. on functions of the form $\Phi(\mathbf{x_i}) \cdot \Phi(\mathbf{x_j})$. Now if there were a kernel function $K$ such that $K(\mathbf{x_i}, \mathbf{x_j}) = \Phi(\mathbf{x_i}) \cdot \Phi(\mathbf{x_j})$ , we would only need to use $K$ in the training algorithm, and would never need to explicitly even know what $\Phi$ is.

An example is a function known as a Radial Basis Kernel:

$$K(\mathbf{x_i}, \mathbf{x_j}) = \exp \frac{-\|\mathbf{x_i} - \mathbf{x_j}\|^2}{2\sigma^2} \tag{4.49}$$

In this particular example, $\Phi(\mathbf{x})$ is infinite dimensional, so it would not be very easy to work with $\Phi$ explicitly. However, if one replaces $\mathbf{x_i} \cdot \mathbf{x_j}$ by $K(\mathbf{x_i} \cdot \mathbf{x_j})$ everywhere in the training algorithm, the algorithm will fortunately produce an SVM which lives in an infinite dimensional space, and furthermore take roughly the same amount of time it would take to train on the unmapped data. All the considerations of the previous sections hold, since linear separation is still performed, but in a different space. In that case, $\mathbf{w}$ that will live in $\Phi(\mathbf{x})$ (see Equation 4.45). However, an SVM is used in test phase by computing dot products of a given test point $\mathbf{x}$ with $\mathbf{w}$, or more specifically by computing the sign of

$$f(\mathbf{x}) = \sum_{m \in S} \alpha_m y_m \Phi(\mathbf{x_m}) \cdot \Phi(\mathbf{x}) + b = \sum_{m \in S} \alpha_m y_m K(\mathbf{x_m}, \mathbf{x}) \tag{4.50}$$

for all Support Vectors in $S$. Therefore, we can avoid computing $\Phi(\mathbf{x})$ explicitly and use the $K(\mathbf{x_m}, \mathbf{x}) = \Phi(\mathbf{x_m}) \cdot \Phi(\mathbf{x_m})$ instead.

When applying the SVM to linearly separable data we have started by creating a matrix $\mathbf{H}$ from the dot product of our input variables:

$$H_{ij} = y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} = \mathbf{x_i}^T \mathbf{x_j} \tag{4.51}$$

where the kernel function $K(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i}^T \mathbf{x_j}$ is known as linear kernel. The set of kernel functions is composed of variants of Equation 4.49 in that they are all based on calculating inner products of two vectors. This means that if the functions can be recast into a higher dimensionality space by some potentially non-linear feature mapping function $\Phi : \mathbf{x} \mapsto \Phi(\mathbf{x})$ , only inner products of the mapped inputs in the feature space need be determined without explicitly calculating $\Phi$. The reason that this Kernel Trick is useful is that there are many classification/regression problems that are not linearly separable/regressable in the space of the inputs x, which might be in a higher dimensionality feature space given a suitable mapping.

Some common kernel functions include:

- Polynomial (homogeneous): $K(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j})^d$
- Polynomial (inhomogeneous): $K(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^d$
- Gaussian Radial Basis Function: $K(\mathbf{x_i}, \mathbf{x_j}) = \exp \frac{-\|\mathbf{x_i} - \mathbf{x_j}\|^2}{2\sigma^2}$
- Hyperbolic tangent: $K(\mathbf{x_i}, \mathbf{x_j}) = \tanh(\kappa \mathbf{x_i} \cdot \mathbf{x_j} + c)$, for some (not every) $\kappa > 0$ and $c < 0$

An example dataset that is not linearly separable in the two dimensional data space becomes separable in the non-linear feature space defined implicitly by non-linear Radial Basis Kernel function (Equation 4.49) shown in Figure 4.5.

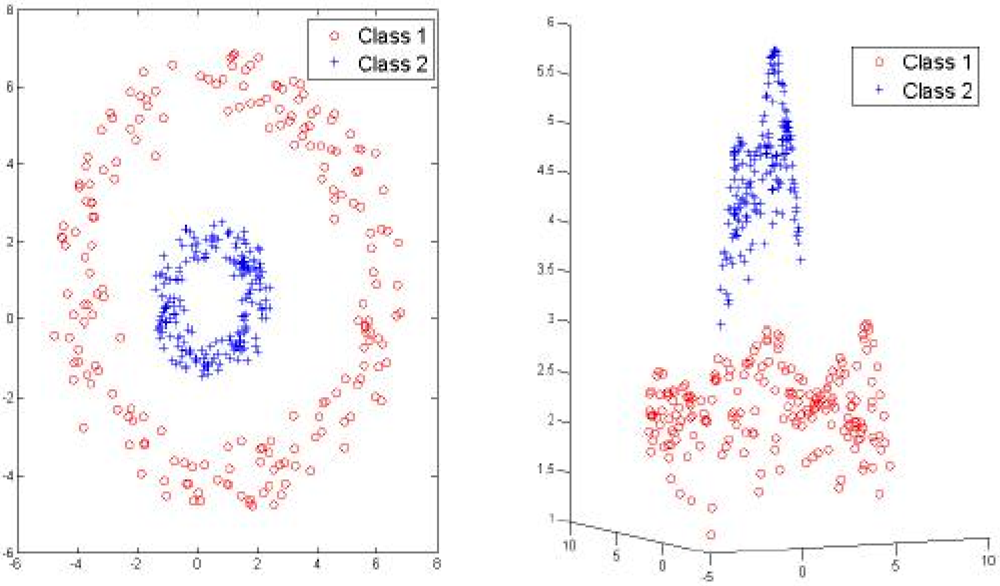


Figure 4.5. Dichotomous data re-mapped using Radial Basis Kernel.

### Parameter selection

The effectiveness of SVM depends on the selection of kernel, the kernel's parameters, and soft margin parameter $C$. A common choice is a Gaussian kernel, which has a single parameter $\sigma$. Best combination of $C$ and $\sigma$ is often selected by a grid-search with exponentially growing sequences of C and $\sigma$, for example, $C \in \{2^{-5}, 2^{-3}, \ldots, 2^{13}, 2^{15}\}$, $1/2\sigma^2 \in \{2^{-15}, 2^{-13}, \ldots, 2^{1}, 2^{3}\}$. Typically, each combination of parameter choices is checked using cross validation, and the parameters with best cross-validation accuracy are picked. The final model, which is used for testing and for classifying new data, is then trained on the whole training set using the selected parameters.

## 4.3.4 A Support Vector Machine Realisation

$SVM^{light}$ is an implementation of Vapnik's Support Vector Machine (Cortes and Vapnik 1995) for the problem of pattern recognition, for the problem of regression, and for the problem of learning a ranking function. The optimization algorithms used in $SVM^{light}$ are described in (Joachims 2002, 1999a). The algorithm has scalable memory requirements and can handle problems with many thousands of Support Vectors efficiently.

The software also provides methods for assessing the generalization performance. It includes two estimation methods for both error rate and precision/recall. By selecting one of the estimation methods, which is known as $XiAlpha$, estimates can be computed at essentially no additional computational expense, but they are conservatively biased (Joachims 2002). Almost unbiased estimates are provided by the leave-one-out cross-validation. $SVM^{light}$ exploits the fact that the results of most leave-one-outs (often more than 99%) are predetermined and need not be computed.

The code has been used on a large range of problems, including text classification (Joachims 1999b, 1998), image recognition tasks, bioinformatics and medical applications. Many tasks have the property of sparse instance vectors. This implementation makes use of this property which leads to a very compact and efficient representation.

$SVM^{light}$ is an implementation of Support Vector Machines (SVMs) in $C$ programming language and available for download in the website (http://svmlight.joachims.org/).

## 4.4 Vector Dimensionality Reduction

Vector dimensionality reduction is a mathematical transformation to represent a vector dataset in fewer dimensions. In a classification problem, the dimensionality reduction is usually performed to improve the classification performance and reduce the computation time.

Researchers are commonly faced with intrinsically low-dimensional structures hidden in very high-dimensional spaces in many areas of science. Finding these meaningful low-dimensional structures from large amounts of data is the problem of dimensionality reduction.

### 4.4.1 The Curse Of Dimensionality

The curse of dimensionality is a term coined in 1961 that refers to the problems associated with multivariate data analysis as the dimensionality increases (Bellman 1961). In practice, the curse of dimensionality means that, for a given sample size, there is a maximum number of features above which the performance of our classifier will degrade rather than improve, as illustrated in Figure 4.6. In most cases, the information lost by discarding some features is compensated by a more accurate mapping in the lower dimensional space.

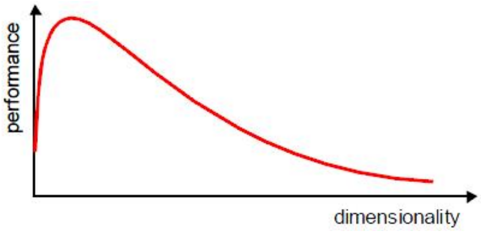


Figure 4.6. The curse of dimensionality for a classifier performance.

### 4.4.2 Feature Selection And Feature Extraction

There are two basic approaches are available for dimensionality reduction:

1. Feature selection; choosing a subset of all the features (the ones that are more informative), see Equation 4.52.

2. Feature extraction; creating a subset of new features via combinations of the existing features, see Equation 4.53.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_p \end{bmatrix} \xrightarrow{feature\,selection} \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{bmatrix} \tag{4.52}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_p \end{bmatrix} \xrightarrow{feature\,extraction} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_p \end{bmatrix}\right) \tag{4.53}$$

The problem of feature extraction can be stated as finding a mapping $f : R^p \to R^m$ with $m < p$ such that the transformed feature vector $f(\mathbf{x}) \in R^m$ preserves (most of) the information or structure in $R^p$.

In general, the optimal mapping $\mathbf{y} = f(\mathbf{x})$ will be a non-linear capturing the manifold that covers the dispersion of the data. On the other hand, there is no systematic way to generate optimal feature-reducing non-linear transforms. In other words, the selection of a particular subset of transforms is problem dependent. For this reason, feature extraction is commonly limited to linear transforms: $\mathbf{y} = \mathbf{W}\mathbf{x}$ in which, $\mathbf{y}$ is a linear projection of $\mathbf{x}$. A very widely used linear approach is metric Multidimensional Scaling (MDS) (Kruskal 1964, Shepard 1962). It aims to represent the data points in a lower dimensional space while preserving as much of the pairwise similarities between the data points as possible.

In our study, we have considered several vector dimensionality reduction methods, namely, Individual Feature Selection, Principal Component Analysis (PCA) (Jolliffe 2002), Isomap (Tenenbaum et al. 2000) as well as Independent Component Analysis (ICA) that are described in the following sections.

### 4.4.3 Feature Selection

The Individual Feature Selection evaluates each texture feature separately. The advantage of individual search is high speed. It is therefore useful for preselection of a candidate feature subset from a large set of features. Note, however, that individually poor features may yield high class separability when used together.

Feature selection is a methodology where a subset of features is selected from the complete feature dataset, according to a defined selection rule or a criterion. The best subset contains the fewest number of dimensions that contribute to the recognition accuracy the most; while the remaining, unimportant dimensions are discarded. The purpose of the feature selection for a classifier is to have better classification performance using a secondary feature vector dataset of lower dimension. Another use of the feature selection is to visualize the data of interest where applicable.

Feature selection methods are grouped as optimal and suboptimal methods (Jain and Zongker 1997);

1. Optimal methods.
   (a) Exhaustive Selection.
   (b) Branch and Bound method.
2. Suboptimal methods.
   (a) Sequential methods.
      i. Forward Selection.
      ii. Backward Selection.
      iii. Variants: Variants of both Forward Selection and Backward Selection.
   (b) Genetic algorithms.

and described as:

Exhaustive Selection should be considered as an idealized feature selection method where all possible feature combinations are formed and the classifier performances evaluated for each feature combination. However, such an exhaustive approach requires examining all possible subsets of a feature set. The number of possibilities grows exponentially, making exhaustive search impractical for even moderate values of dimensionality. Therefore, Exhaustive Selection is generally not applicable for many classifiers as in our texture classification problem since the number of texture feature combinations and the total processing time are considerably high.

The Branch and Bound (BB) feature selection algorithm can be used to find the optimal subset of features much more quickly than an exhaustive search (Narendra and Fukunaga 1977). One drawback is that the Branch and Bound procedure requires the feature selection criterion function to be monotone. This means that the addition of new features to a feature subset can never decrease the value of the criterion function. On the other hand, as is known from the curse of dimensionality phenomenon that in small sample size situations this may not be true. Furthermore, the Branch and Bound method is still impractical for problems with very large feature sets, because the worst case complexity of this algorithm is exponential.

The sequential methods begin with a single solution (a feature subset) and iteratively add or remove features until some termination criteria is met. These are the most commonly used methods for feature selection. They can be divided into two categories, those that start with the empty set and add features (the forward methods) and those that start with the full set and delete features (the backward methods). Note that since they don't examine all possible subsets, these algorithms are not guaranteed to produce the optimal result. Indeed, it was shown that no non-exhaustive sequential feature selection procedure can be guaranteed to produce the optimal subset (Cover and Campenhout 1977).

A Forward Selection algorithm starts with no features and add them one by one, at each step adding the one that decreases the error the most, until any further addition does not significantly decrease the error. On the other hand, a Backward Selection algorithm starts with all the variables and remove them one by one, at each step removing the one that decreases the error the most (or increases it only slightly), until any further removal increases the error significantly. The sequential algorithms known as Variant Algorithms are the mixtures of both Forward

Selection and Backward Selection algorithms that at each step, consider all additions and removals of each feature, and select the best combination. Kittler published a comparative study of these algorithms and the optimal branch-and-bound algorithm using a synthetic two-class Gaussian dataset (Kittler 1978).

Genetic Algorithms (GA) were introduced for feature selection by (Siedlecki and Sklansky 1989). In a GA approach, a given feature subset is represented as a binary string (chromosome) of length $p$, with a zero or one in position $i$ denoting the absence or presence of feature $i$ in the set. Note that $p$ represents the total number of available features. Each chromosome is evaluated to determine its fitness which determines how likely the chromosome is to survive and breed into the next generation. New chromosomes are created from old chromosomes by the processes of:

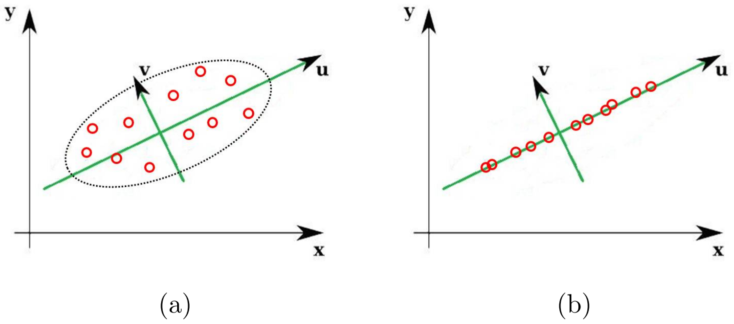
1. Crossover; where parts of two different parent chromosomes are mixed to create offspring,

2. Mutation; where the bits of a single parent are randomly perturbed to create a child. The chromosome that survive after many generations then represent the feature combinations that produce high classification performance.

### 4.4.4 Principal Component Analysis

The Principal Component Analysis (PCA) is a statistical multivariate data analysis method that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a linearly uncorrelated set of variables called principal components. This representation can be considered as the transformation of the original data into a new vector space where the basis vectors are actually a linear combination of the original data vectors. PCA can be briefly described as the projection of the multivariate data on the eigenvectors of the covariance matrix of the original data (Jolliffe 2002). The amount of variance preserved by the projected data in a certain principal component (eigenvector) direction is given by the eigenvalue corresponding to that direction.

Suppose that we have a dataset measured in the $x-y$ coordinate system as shown in Figure 4.7(a). The principal direction in which the data varies is shown by the $u$ axis and the second most important direction is the $v$ axis orthogonal to it. If the $u-v$ axes are placed at the mean of the data it gives us a compact representation. If each (x,y) coordinate is transformed into its corresponding $(u, v)$ value, the data becomes de-correlated, meaning that the correlation between the $u$ and $v$ variables is zero. Thus, for a given dataset, PCA finds the axis system defined by the principal directions of variance. In this example, the directions $u$ and $v$ are called the principal components.



(a) (b)

Figure 4.7. PCA for (a) data representation, (b) dimensionality reduction.

Let us also consider how PCA offers a way of reducing the dimensionality of a dataset. Consider two variables that have an almost linear relation as shown in Figure 4.7(b). As in Figure 4.7(a) the principal direction in which the data varies is shown by the $u$ axis, and the secondary direction by the $v$ axis. However, in this case all the $v$ coordinate values are all spread around zero. Therefore, it may be assumed that they are caused by noise. Thus, in the $u-v$ axis system, the data set can be represented by one variable $u$ discarding the variable $v$, reducing the dimensionality of the problem by 1.

Generalising this example, PCA dimensionality reduction can be defined as the optimal approximation of a random vector $\mathbf{x} \in R^p$ by a linear combination of $M$ independent vectors with $M < p$ that is obtained by projecting the random vector $x$ onto the eigenvectors $\Phi_i$ corresponding to the largest eigenvalues $\lambda_i$ of the covariance matrix.

In the implementation of the PCA method, since the variance depends on the scale of the variables, it is customary to first standardize each variable to have mean zero and standard deviation one. After the standardization, the original variables with possibly different units of measurement are all in comparable units. Assuming a standardized data with the empirical covariance matrix

$$\mathbf{\Sigma_{pxp}} = \frac{1}{n}\mathbf{X}\ \mathbf{X}^T \tag{4.54}$$

assuming there are $n$ observations and $\mathbf{x}$ is a $p$ dimensional random variable where $\mathbf{x} = (x_1, \ldots, x_p)^T$ and the observation matrix is $\mathbf{X} = \{x_{i,j} : 1 \leq i \leq p, 1 \leq j \leq n\}$. We can use the spectral decomposition theorem to write covariance matrix $\Sigma$ as

$$\mathbf{\Sigma} = \mathbf{U\Lambda U}^T \tag{4.55}$$

where $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$ is the diagonal matrix of the ordered eigenvalues $\lambda_1 \leq \ldots \leq \lambda_p$ and $\mathbf{U}$ is a $p \times p$ orthogonal matrix containing the eigenvectors. It can be shown that the principal components are given by the $p$ rows of the $p \times n$ matrix $\mathbf{S}$, where (Mardia et al. 1995)

$$\mathbf{S} = \mathbf{U}^T\mathbf{X} \tag{4.56}$$

### 4.4.5 Isomap And Locally Linear Embedding

In this section, two core distance-preserving methods, Isomap and Locally Linear Embedding are summarized. In addition, the Isomap algorithm is presented in detail since this algorithm was used in the experiments presented in Section 5.7.3.

It should be noted that the real-world data usually incorporates non-linear structures rather than linear compositions. Taking this into account, two promising distance-preserving methods, Isomap and Locally Linear Embedding, have been proposed and successfully applied (Friedrich 2002, Roweis and Saul 2000, Tenenbaum et al. 2000). The Isomap algorithm extends MDS by a sophisticated distance measurement to achieve non-linear embeddings. It builds a graph on the data consisting of local connections only, and then measures pairwise distances by the length of the shortest path on that graph. This computes an approximation to the distance between points measured within the underlying manifold. Finally, MDS is used to find a set of low-dimensional points with similar pairwise distances.

The Locally Linear Embedding (LLE) algorithm computes the best coefficients to approximate each point by a weighted linear combination of its neighbours, and then try to find a set of low-dimensional points, which can be linearly approximated by its neighbours with the coefficients determined from the high-dimensional points (Roweis and Saul 2000).

Both these core algorithms are simple to implement, have a very small number of free parameters, and do not get trapped in local minima like many other popular learning algorithms. Furthermore, both have been shown to yield impressive results on artificial and real datasets in comparison to some other non-linear methods such as the Self-Organizing Map (SOM) and Generative Topographic Mapping (GTM) (Bishop et al. 1998, Kohonen 1990).

When reducing the dimensionality of a high-dimensional dataset using these algorithms, only the local neighbourhood structure between the data points remains. This means that Euclidean distances are only meaningful between nearby points. LLE exploits this by describing each point only by its neighbours and finding the best neighbourhood-preserving lower-dimensional representation. On the other hand, Isomap measures the distance on the manifold and tries to obtain a lower-dimensional embedding with these approximated geodesic distances. Figure 4.8 llustrates the Euclidean and the geodesic distance metrics. The difference between Euclidean and geodesic distance is exemplified by two points in a spiral (Lee et al. 2002). The spiral is embedded in a two-dimensional space, but clearly its intrinsic dimension is only one, because one parameter suffices to describe the spiral (Fukunaga 1982). The Euclidean distance in (a) in the higher-dimensional space does not reflect the intrinsic similarity of the two points, as measured by the geodesic distance in (b) along the manifold.

The Isomap algorithm flow is as follows.

1. Firstly, the neighbourhood for each point is calculated in the original high dimensional space. The neighbourhood of a point may be either the $k$ nearest points or the set of points within a radius of $\epsilon$.

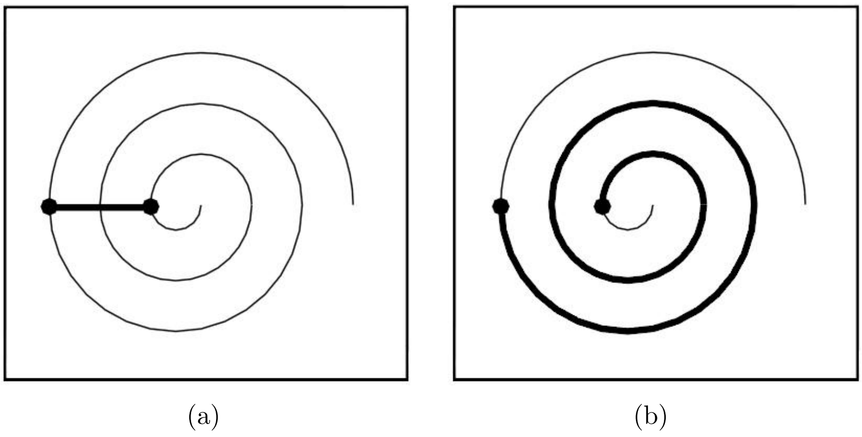


Figure 4.8. Difference between Euclidean and geodesic distance, (a) Euclidean distance, (b) Geodesic distance.

2. After the neighbourhoods are determined, a graph is constructed by linking all neighbouring points and labelling all arcs with the Euclidean distance between the corresponding linked points.

3. The geodesic distance between any two points is approximated by the sum of the arc lengths along the shortest path linking both points. To compute the shortest paths, a more efficient algorithm was suggested that exploits the sparse structure of the neighbourhood graph, presented in (Kumar et al. 1993).

4. Finally classical metric MDS is applied on the approximated geodesic distance matrix, i.e. their largest eigenvectors are computed. The eigenvectors give the coordinates of the data points in the lower-dimensional projection space.

### 4.4.6 Independent Component Analysis

The Independent Component Analysis (ICA) is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals. In ICA, the goal is to find a linear representation of non-Gaussian data so that the components are statistically independent, or as independent as possible. Such a representation captures the essential structure of the data in many applications, including feature extraction and signal separation.

ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples. In the model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed to be non-Gaussian for identifiability purposes and mutually independent, and they are referred to as the independent components of the observed data. These independent components, also called sources or factors can be determined by ICA.

ICA can also be viewed as a vector dimensional reduction approach that seeks to express a multi-variate distribution as a linear combination of statistically independent non-Gaussian random variables (Comon 1994, Hyvärinen et al. 2001), (http://www.cs.helsinki.fi/u/ahyvarin/whatisica.shtml). The ICA method is widely used in the areas of signal source separation and feature extraction.

The technical challenge behind ICA can be expressed by the famous *cocktail-party problem*; imagine that you are in a room where two people are speaking simultaneously and two microphones are recording the time signals at different locations (Haykin and Chen 2005). The recordings can be defined as the time signals $x_1(t)$ and $x_2(t)$ and each of these recorded signals is a weighted sum of the speech signals emitted by the two speakers, which are denoted by $s_1(t)$ and $s_2(t)$. These mixed speech signals can be formulated as

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) \tag{4.57}$$

$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) \tag{4.58}$$

where $a_{11}$, $a_{12}$, $a_{21}$, and $a_{22}$ are parameters that are related to the distances of the microphones to the speakers. The goal is to estimate the two original speech signals $s_1(t)$ and $s_2(t)$, using only the recorded signals $x_1(t)$ and $x_2(t)$.

In a more general setting, we can observe linear mixtures of $n$ independent components where

$$x_j = a_{j1}s_1 + a_{j2}s_2 + ... + a_{jn}s_n \, for \; j = 1, \ldots, n \tag{4.59}$$

in which the time index $t$ is now dropped in the model, therefore, each mixture $x_j$ as well as each independent component $s_k$ is assumed as a random variable, instead of a proper time signal.

Let $\mathbf{x}$ denote the random vector whose elements are the mixtures $x_1, \ldots, x_n$ and likewise $\mathbf{s}$ denote the random vector with elements $s_1, \ldots, s_n$. Let also $\mathbf{A}$ be the matrix of mixture elements $a_{ij}$. Using the matrix notation, Equation 4.59 can be revised as

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{4.60}$$

This ICA model is a generative model, which means that it describes how the observed data are generated by mixing the components $s_i$. Since the independent components cannot be directly observed, they are latent variables. Also the mixing matrix $\mathbf{A}$ is assumed to be unknown. All that is available is the random vector $\mathbf{x}$, and both the sources $s$ are to be estimated using it. Therefore, after estimating the de-mixing matrix $\mathbf{W}$, the sources can be computed simply by:

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{4.61}$$

### 4.4.7 Principles Of The Independent Component Analysis

In this section, the principles of the Independent Component Analysis are described.

### Independence Of Non-Gaussian Distributions

Basically, ICA model estimation is based on the mixing variables having non-Gaussian distributions. This is the basic system identifiability condition on the ICA problem (Hyvärinen et al. 2001). The distribution of a sum of independent random variables tends toward a Gaussian distribution, under certain conditions by the Central Limit Theorem. Thus, the sum of two independent random variables usually can be expected to have a distribution that is closer to a Gaussian than the two original random variables.

## Measures Of Non-Gaussianity

To use non-Gaussianity in ICA estimation, a quantitative measure of non-Gaussianity of a random variable, say $y$, should be devised. Several non-Gaussianity measures are detailed below.

### Kurtosis

Kurtosis can be described as the degree of peakedness of a distribution. Kurtosis is the fourth-order classical measure of non-Gaussianity defined by

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2 \tag{4.62}$$

Note that, for a Gaussian random variable $y$, the fourth moment equals to $3(E\{y^2\})^2$, and hence, the kurtosis is zero.

Typically non-Gaussianity is measured by the absolute value of kurtosis. The square of kurtosis can also be used. These measures tend to zero for a Gaussian variable, and greater than zero for most non-Gaussian random variables. There are non-Gaussian random variables that have zero kurtosis, but they can be considered very rare.

On the other hand, kurtosis has also some drawbacks in practice, when its value has to be estimated from a measured sample. The main problem is that kurtosis can be very sensitive to outliers (Huber 1985). In that case, kurtosis values may depend on only a few marginal observations in the tails of the distribution, which may generate erroneous or irrelevant values. In other words, kurtosis is not a robust measure of non-Gaussianity. Thus, other measures of non-Gaussianity might be better than kurtosis in some situations.

### Negentropy

Negentropy is considered to be a better non-Gaussianity measure than kurtosis, whose properties are rather opposite to those of kurtosis. Entropy is the one of the basic concepts of information theory. The entropy of a random variable can be interpreted as the degree of information that the observation of the variable provides. The more unpredictable and unstructured the variable is, the larger its entropy. The entropy $H$ of a random variable $Y$ is defined as

$$H(Y) = -\sum_i P(Y = a_i)\log(P(Y = a_i)) \tag{4.63}$$

where the $a_i$ are the possible values of $Y$ (Cover and Thomas 1991, Papoulis 1991). This very well-known definition can be generalized for continuous-valued random variables and vectors, in which case it is often called differential entropy. The differential entropy $H$ of a random vector $\mathbf{y}$ with density $f(\mathbf{y})$ is defined as (Cover and Thomas 1991, Papoulis 1991)

$$H(\mathbf{y}) = -\int f(\mathbf{y})\log(f(\mathbf{y}))dy \tag{4.64}$$

A fundamental result of information theory is that a Gaussian variable has the largest entropy among all random variables of equal variance (Cover and Thomas 1991, Papoulis 1991). This means that entropy could be used as a measure of non-Gaussianity. In fact, this shows that the Gaussian distribution is the "most random" or the least structured of all distributions.

To obtain a measure of non-Gaussianity, a slightly modified version of the definition of differential entropy, called negentropy, is used. Negentropy $J$ is defined as

$$J(\mathbf{y}) = H(\mathbf{y_{gauss}}) - H(\mathbf{y}) \tag{4.65}$$

where $\mathbf{y_{gauss}}$ is a Gaussian random variable of the same covariance matrix as $\mathbf{y}$. Due to the above-mentioned properties, negentropy is always non-negative, and it is zero if and only if $\mathbf{y}$ has a Gaussian distribution.

The main disadvantage of negentropy is the computational cost, because the computation of negentropy requires the estimation of the probability density function. For some practical approximations to this, please see (Hyvärinen and Oja 2000).

## Minimization Of Mutual Information

Another approach for ICA estimation, inspired again by information theory, is minimization of mutual information. This approach leads essentially to the same principle of finding he most non-Gaussian sources as was described above.

Using the concept of differential entropy, the mutual information $I$ between $m$ random variables, $y_i$ for $i = 1, \ldots, m$ is defined as follows:

$$I(y_1, y_2, ..., y_m) = \sum_i^m H(y_i) - H(\mathbf{y}) \tag{4.66}$$

Mutual information is a natural measure of the dependence between random variables. It is always non-negative, and zero if and only if the variables are statistically independent. Thus, the mutual information takes into account the whole dependence structure of the variables, and not only the covariance, like PCA and the other related methods.

An important property of mutual information for an invertible linear transformation is given by $\mathbf{y} = \mathbf{W}\mathbf{x}$ (Papoulis 1991, Cover and Thomas 1991) :

$$I(y_1, y_2, ..., y_m) = \sum_i^m H(y_i) - H(\mathbf{x}) - \log \mid \det(\mathbf{W}) \mid \tag{4.67}$$

Considering that $y_i$ are constrained to be uncorrelated and of unit variance provides $E\{\mathbf{y}\mathbf{y}^T\} = \mathbf{W}E\{\mathbf{x}\mathbf{x}^T\}\mathbf{W}^T = \mathbf{I}$, which implies

$$\det \mathbf{I} = 1 = \det(\mathbf{W}E\{\mathbf{x}\mathbf{x}^T\}\mathbf{W}^T) = (\det \mathbf{W})(\det E\{\mathbf{x}\mathbf{x}^T\})(\det \mathbf{W}^T), \tag{4.68}$$

and this also implies that $det\mathbf{W}$ must be constant. Moreover, for $y_i$ of unit variance, entropy and negentropy differ only by a constant, and the sign. Thus,

$$I(y_1, y_2, ..., y_m) = C - \sum_i (J(y_i)) \tag{4.69}$$

where $C$ is a constant that is independent of $\mathbf{W}$. This equation shows the fundamental relation between the negentropy and the mutual information.

Since mutual information is the natural information-theoretic measure of the independence of random variables, it can be used as a criterion for finding the ICA transform. In an alternative approach, the ICA of a random vector $\mathbf{x}$ can be defined as an invertible transformation as in Equation 4.61, where the matrix $\mathbf{W}$ is determined so that the mutual information of the transformed components $s_i$ is minimized. It is now obvious from Equation 4.69 that finding an invertible transformation $\mathbf{W}$ that minimizes the mutual information is roughly equivalent to finding directions in which the negentropy is maximized. Rigorously speaking, this shows that ICA estimation by minimization of mutual information is equivalent to maximizing the sum of non-Gaussianities of the estimates, when the estimates are constrained to be uncorrelated. The constraint of uncorrelatedness is in fact not necessary, but simplifies the computations considerably, as one can then use the simpler form in Equation 4.69 instead of the more complicated form in Equation 4.67.

It can also shown that maximum likelihood estimation is essentially equivalent to minimization of mutual information and can be used in estimating the suitable ICA model.

It is possible to formulate directly the likelihood in the noise-free ICA model, as was done by (Pham et al. 1992), and then estimate the model by a maximum likelihood method. Denoting the matrix $\mathbf{A}^{-1}$ by $\mathbf{W} = (w_1, \ldots, w_n)^T$, the log-likelihood takes the form:

$$L = \sum_{t=1}^{T} \sum_{i=1}^{n} \log f_i(w_i^T x(t)) + T \log \mid \det(\mathbf{W}) \mid \tag{4.70}$$

where the $f_i$ are the density functions of the $s_i$ (here assumed to be known), and the $x(t), t = 1, \ldots, T$ are the realizations of $\mathbf{x}$ (Papoulis 1991).

## Preprocessing For ICA

Before applying an ICA algorithm to the data, it is usually very useful to do some preprocessing. In this section, some preprocessing techniques, namely, centering and whitening, that make the problem of ICA estimation simpler and better conditioned are discussed.

## Centering

The most basic and necessary preprocessing is to center $\mathbf{x}$, i.e. to subtract its mean vector $\mathbf{m} = E\{\mathbf{x}\}$ so as to make $\mathbf{x}$ a zero-mean variable. This implies that $s$ is zero-mean as well, which can be seen by taking expectations of both sides of Equation 4.60.

This preprocessing is made solely to simplify the ICA algorithms. After estimating the mixing matrix $\mathbf{A}$ with centered data, one can complete the estimation by adding the mean vector of s back to the centered estimates of $s$. The mean vector of $s$ is given by $\mathbf{A}^{-1}\mathbf{m}$, where $\mathbf{m}$ is the mean that was subtracted in the preprocessing.

## Whitening

Another useful preprocessing strategy in ICA is first to whiten the observed variables after centering. This means that before the application of the ICA algorithm, the centered vector $\mathbf{x}$ is transformed linearly so that a new vector $\tilde{\mathbf{x}}$ is obtained that is white, hence, its components are uncorrelated and their variances equal to unity. In other words, the covariance matrix of $\tilde{\mathbf{x}}$ equals the identity matrix, expressed by

$$E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathbf{T}}\} = \mathbf{I} \tag{4.71}$$

It should be noted that the whitening reduces the number of parameters to be estimated. Instead of having to estimate the $n^2$ parameters that are the elements of the original matrix $\mathbf{A}$, we only need to estimate the new, orthogonal mixing matrix $\tilde{A}$. Note that, an orthogonal matrix contains $n(n-1)/2$ degrees of freedom. Therefore, whitening essentially solves half of the problem. Because the whitening is a very simple and standard procedure, much simpler than any of the ICA algorithms, it is a very good idea to reduce the complexity of the problem.

### 4.4.8 The FastICA Algorithm

The FastICA algorithm (http://research.ics.tkk.fi/ica/fastica) is a popular ICA method that uses a fixed point iteration scheme to maximise the non-Gaussianity of the unknown sources (Hyvärinen and Oja 1997, Hyvärinen 1999). In independent experiments, FastICA has been found to be 10-100 times faster than conventional gradient descent based ICA methods.

In the preceding sections, different measures of non-Gaussianity, i.e. objective functions for ICA estimation or contrast functions as commonly referred to were introduced. In practice, an algorithm for maximizing the contrast function is also needed. In this section, the very efficient FastICA method is introduced.

As the preliminary part of the FastICA, it is assumed that the data first has been preprocessed by centering and whitening as explained above. For simplicity of notation, we denote the preprocessed data just by $\mathbf{x}$, and the transformed mixing matrix by $\mathbf{A}$, omitting the tildes.

The FastICA learning rule is to find a direction, i.e. a unit vector $\mathbf{w}$ such that the projection $\mathbf{w}^T\mathbf{x}$ maximizes non-Gaussianity. Non-Gaussianity is measured by the approximation of negentropy $J(\mathbf{w}^T\mathbf{x})$. Recall that the variance of $\mathbf{w}^T\mathbf{x}$ must here be constrained to unity; for whitened data this is equivalent to constraining the norm of $\mathbf{w}$ to be unity. The FastICA is based on a fixed-point iteration scheme for finding a maximum of the non-Gaussianity of $\mathbf{w}^T\mathbf{x}$, see (Hyvärinen and Oja 1997, Hyvärinen 1999). It can be also derived as a Newton iteration. Let $g$ be the derivative of a non-quadratic non-linearity; valid choices are

$$g_1(u) = \tanh(a_1 u),\ g_2(u) = u\exp(-u^2/2) \tag{4.72}$$

where the corresponding non-quadratic functions are

$$G_1(u) = \frac{1}{a_1}\log(\cosh(a_1 u)),\ G_2(u) = -\exp(u^2/2) \tag{4.73}$$

where $1 \leq a_1 \leq 2$ is some suitable constant, often taken as $a_1 = 1$. The basic form of the FastICA algorithm is as follows (Hyvärinen and Oja 2000):

1. Choose an initial (e.g. random) weight vector $\mathbf{w}$.

2. Let $\mathbf{w}^+ = E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}$

3. Let $\mathbf{w}^+ = \mathbf{w}^+/ \parallel \mathbf{w}^+ \parallel$

4. If not converged, go back to Step 2.

The convergence is achieved when the old and new values of $\mathbf{w}$ point in the same direction, i.e. their dot-product is almost equal to 1. It is not necessary that the vector converges to a single point, since $\mathbf{w}$ and $-\mathbf{w}$ define the same direction.

The derivation of FastICA is as follows. First note that the maxima of the approximation of the negentropy of $\mathbf{w}^T\mathbf{x}$ are obtained at certain optima of $E\{G(\mathbf{w}^T\mathbf{x})\}$. According to the Kuhn-Tucker conditions (Luenberger 1969), the optima of $E\{G(\mathbf{w}^T\mathbf{x})\}$ under the constraint $E\{(\mathbf{w}^T\mathbf{x})^2\} = \parallel \mathbf{w} \parallel^2 = 1$ is obtained at points where

$$E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - \beta\mathbf{w} = 0 \tag{4.74}$$

Denoting the function on the left-hand side of Equation 4.74 by $F$ and solving this equation by the Newton's method, its Jacobian matrix $JF(\mathbf{w})$ is obtained as

$$JF(\mathbf{w}) = E\{\mathbf{x}\mathbf{x}^T g'(\mathbf{w}^T\mathbf{x})\} - \beta\mathbf{I} \tag{4.75}$$

To simplify the inversion of this matrix, it is decided to approximate the first term in Equation 4.75. Since the data is sphered, a reasonable approximation is given by

$$E\{\mathbf{x}\mathbf{x}^T g'(\mathbf{w}^T\mathbf{x})\} \approx E\{\mathbf{x}\mathbf{x}^T\}E\{g'(\mathbf{w}^T\mathbf{x})\} = E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{I} \tag{4.76}$$

Thus the Jacobian matrix becomes diagonal, and can easily be inverted. Then, the following approximative Newton iteration is obtained :

$$\mathbf{w}^{+} = \mathbf{w} - (E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - \beta\mathbf{w})/(E\{g'(\mathbf{w}^T\mathbf{x})\} - \beta) \tag{4.77}$$

This algorithm can be further simplified by multiplying both sides of Equation 4.77 by $(E\{g'(\mathbf{w}^T\mathbf{x})\} - \beta)$ which gives, after algebraic simplification, the FastICA iteration (Hyvärinen and Oja 2000).

## Properties of the FastICA Algorithm

The FastICA algorithm and the underlying contrast functions have a number of desirable properties when compared with existing methods for ICA (Hyvärinen and Oja 2000);

1. The convergence is cubic (or at least quadratic), under the assumption of the ICA data model (Hyvärinen 1999). This is in contrast to ordinary ICA algorithms based on (stochastic) gradient descent methods, where the convergence is only linear. This means a very fast convergence.

2. FastICA algorithm is easy to use contrary to gradient-based algorithms since there is no step size parameter to choose.

3. The algorithm finds directly independent components of (practically) any non-Gaussian distribution using any non-linearity $g$. This is in contrast to many algorithms, where some estimate of the probability distribution function has to be first available, and the non-linearity must be chosen accordingly.

4. The performance of the method can be optimized by choosing a suitable non-linearity $g$. In particular, one can obtain algorithms that are robust and/or of minimum variance.

5. The independent components can be estimated one by one, which is roughly equivalent to doing projection pursuit. This is useful in exploratory data analysis, and decreases the computational load of the method in cases where only some of the independent components need to be estimated.

6. The FastICA has most of the advantages of neural algorithms: It is parallel, distributed, computationally simple, and requires little memory space. Stochastic gradient methods seem to be preferable only if fast adaptivity in a changing environment is required.

## 4.5 Proposed Framework

In the above sections of this chapter, the detailed background information on the elementary blocks of the proposed framework was presented. Now, the main framework integrating these blocks is summarized. The proposed framework is described in the form of the flow diagrams in Figure 4.9. The QSL texture labelling and the QSL target texture classification methodology are presented separately.

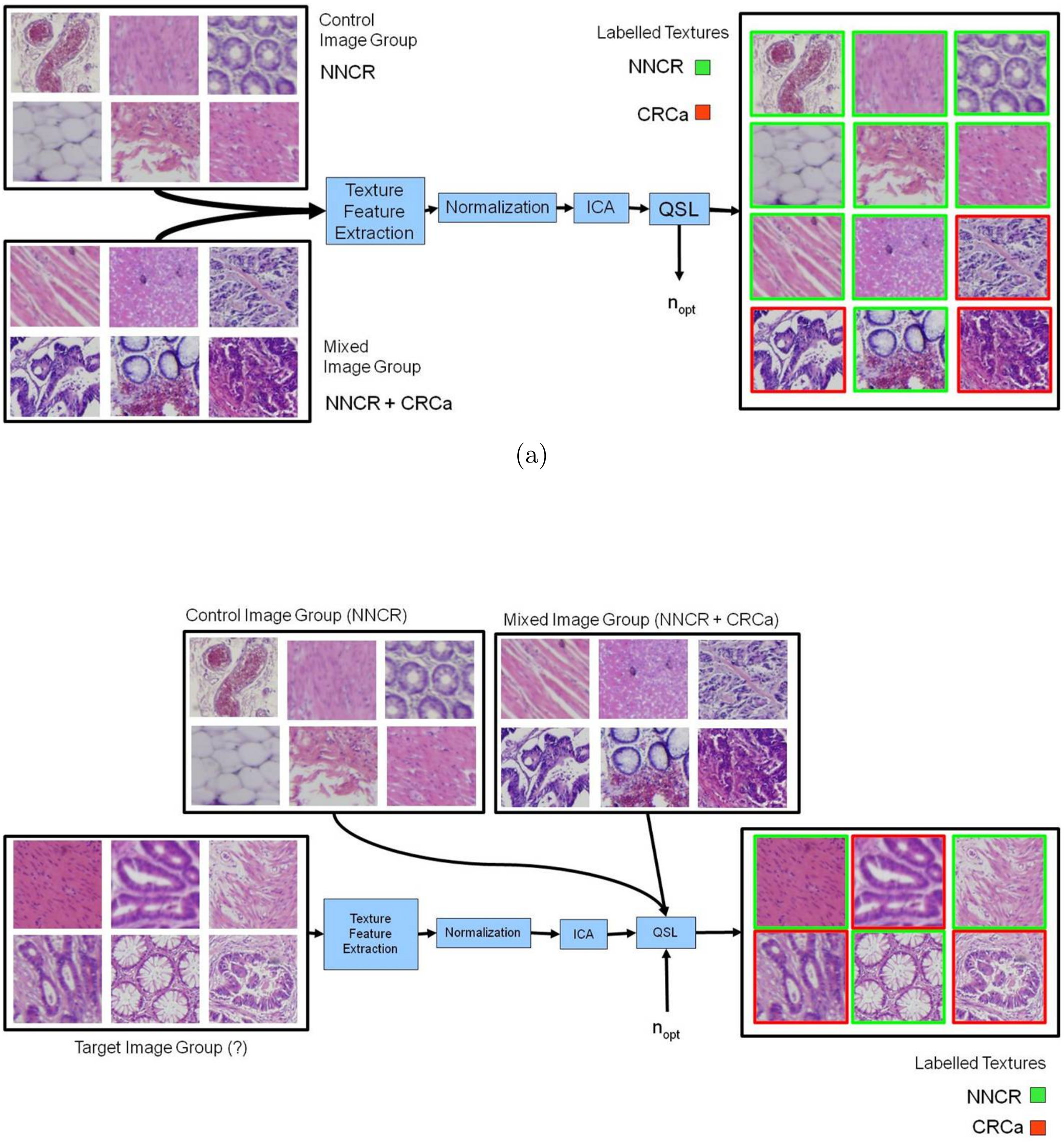
### 4.5.1 Texture Labelling By The Quasi-supervised Learning

The first block of the proposed framework is texture feature extraction as explained in Section 4.1. In this block, the texture feature vectors are calculated using the histopathology images from the reference and the mixed image groups . Each feature vector takes the initial label of its source image group; either reference or mixed.

The block following the texture feature extraction is the vector dimensionality reduction block (for the background see Section 4.4). As corresponding experimental results will be presented in Section 5.7, the only dimensionality reduction method that improved the texture labelling performance was the ICA procedure.

Following the dimensionality reduction, all of the reduced feature vectors are then fed to the QSL algorithm. The QSL algorithm calculates the posterior probability $p_0(\mathbf{x_i})$ for each texture feature vector $x_i$, defined as the probability of it being assigned to the reference label. At the same time, the optimum reference set size $n_{opt}$ for these texture vector datasets are also calculated (see Section 4.2.3).

After $p_0(\mathbf{x})$ values are calculated for all feature vectors $x_i$, the threshold $T_{opt}$ is to be determined in order to set resultant texture vector labels. The feature



(a)

(b)

Figure 4.9. Graphical abstract of the proposed framework (a) QSL texture labelling, (b) QSL target texture classification.

vectors $\mathbf{x}$ that satisfy the condition $p_0(\mathbf{x}) \geq T_{opt}$ are assigned to the label reference, whereas, all other vectors are assigned to the label of contrast. The meaning of the contrasted label may vary due to the experimental setup but it is basically the contrasted characteristics between reference and unlabelled mixed image groups. In our study, the reference image label corresponds to the healthy regions (NNCR) whereas contrasted label, i.e. the contrast between mixed and the reference groups, means cancer (CRCa) or any other unknown local characteristics specific to mixed image group.

To display the labelling results, all local image regions are framed as an overlay onto the original histopathology images by using two different colors. The green color represents textures labelled as healthy and the red color represents textures labelled as cancer-related.

### 4.5.2 Classification Of Target Textures By The Quasi-supervised Learning

The methodology that was graphically abstracted in Figure 4.9(a) is an analytical use of the QSL algorithm for texture labelling. In addition to the analytical use, the QSL algorithm was also extended to classify the texture feature vectors observed after the original dataset. This is the predictive use of the QSL algorithm which we call target texture classification. The flow diagram of the proposed predictive methodology is shown in Figure 4.9(b). This diagram is very similar to the first flow diagram except for the additional target images at the vector data input part. Another difference is the $n_{opt}$ input that is to be identified as a result of the analytical learning on the original texture feature vector.

In the proposed method all texture feature vectors were calculated for all image regions in the histopathology images from the reference, mixed and the target image groups. The feature vectors are then fed to the QSL algorithm that calculates the posterior probability $p_0(x)$ for each target feature vector $\mathbf{x}$ against the reference and mixed vectors using the provided optimum set size $n_{opt}$.

After the $p_0(\mathbf{x})$ values are determined for each target feature vector $\mathbf{x}$, a threshold $T_{opt}$ is determined in a fashion that is similar to the texture labelling case. Finally, the target feature vectors $\mathbf{x}$ that satisfy the condition $p_0(\mathbf{x}) \geq T_{opt}$ are assigned to the reference label and all other target vectors are assigned to the label of contrast.

The labels of the image regions are then displayed as before, overlaid on the color histopathology images.

# CHAPTER 5

# EXPERIMENT SETUP AND RESULTS

This chapter presents the setup of our study along with the results of the experimental execution. Firstly, the histopathological digital image database of the study, the Histopathological Image Library (HIL) is introduced. Secondly, the preferred approach for the performance evaluation of the QSL labelling and classification is explained together with the ground truth data extraction tool. Following the performance evaluation, the texture information base used throughout this study which are texture feature vector datasets calculated using various extraction parameters are detailed. Then, the resultant labelling performances of the QSL procedure applied to the reference and the mixed vector groups were presented along with the corresponding experimental details. Next, the effects of varying texture extraction parameters to the labelling performance were investigated. The study performed in order to have performance comparative information of the QSL method against a powerful supervised classifier is explained in Section 5.6. Next, the studies performed to determine the QSL algorithm response to vector dimensionality reduction are presented. The following two sections are dedicated to synthetic validation of the QSL labelling and to background segmentation techniques. The last sections are concerned with the predictive use of the QSL algorithm on the feature vectors of unknown origin.

## 5.1 Histopathological Image Library

In this study, the Histopathological Image Library (HIL), consisting of 257 light microscopic images was constructed to be used in the texture classification experiments. This set of digital light microscopic images was taken from hematoxylin and eosin (H&E) stained sections of formalin-fixed paraffin embedded tissue sections of non-neoplastic colorectal (NNCR) and colorectal carcinoma (CRCa) tissues from radical colectomy or rectum resection specimens by a digital camera (Olympus

DP70, Olympus Optical Co. Ltd., Tokyo, Japan) connected to a light microscope (Olympus BX51, Olympus Optical Co. Ltd., Tokyo, Japan) at an original magnification of ×4.

All images in the library have 256 level $RGB$ planes, $4080 \times 3720$ pixel dimensions and approximately 1 *micron* per pixel resolution. The images were acquired using fixed capture and illumination parameters; (the microscope light exposure was manually set to 6 from a scale of 0 to 6). In this thesis study, $4 \times 4$ pixels regular grid sampling was applied to all original primary images to get smaller size that yielded relatively shorter computation times in feature extraction stage. Thus, the secondary histopathological images used throughout this research had 4 *micron* per pixel resolution.

The images in the HIL were divided into two groups, namely, the training (as reference and mixed) and target image groups. The training image set was used in the QSL texture labelling performance evaluation. On the other hand, the target image set were never included in the QSL training phase, but only queried against the training set in order to measure the performance of the classification. The 27 of the images were assigned as target images while the remaining 230 images were assigned as training images.

The training image group of the library was also divided into two groups, as NNCR and unlabelled mixed; the latter possessing an unlabelled mixture of NNCR and CRCa images. The rule for group assignment was as follows; when an expert observed no carcinoma regions throughout an image, that image was labelled as NNCR and assigned to the first group (NNCR). For the second group, the term "mixed" is used in order to indicate that these images are composed of features associated with both NNCR and CRCa tissues. This separation of the image data into two groups by a pathology expert was a very simple task compared to manual labelling of isolated colorectal carcinoma regions and it was supposed to be much less operator dependent task. The NNCR group had 127 images and the mixed group had 103 images in total.

## 5.2 Texture Labelling Performance Evaluation

This section describes the strategy that uses ground truth texture information in order to evaluate the labelling performance of the proposed framework.

### 5.2.1 Histopathological Image Atlas

Texture feature vector datasets and corresponding group labels as reference or unlabelled mixed were fed to the QSL algorithm and the resulting labels compared with the ground-truth vector label data available in the Histopathological Image Atlas to measure the labelling performance. In order to obtain the ground-truth atlas data, a software tool, the Histopathological Image Atlas Editor (HIAE) was developed in C++ with a Graphical User Interface for the Windows operating system (Microsoft Corporation, USA). The HIAE retrieves selected images from the Histopathological Image Library and allows an expert to mark the cancer regions by mouse. Each image in the library was divided into a grid of 128.0 *microns* and the labelling was done for each square block manually by a pathology expert using the HIAE. This data constitutes the ground-truth data used to evaluate the performance of the proposed method.

A screen snapshot of the HIAE is presented in Figure 5.1. In this figure, the main window of the HIAE displaying an histopathology slide under examination is presented. Individual square regions that were marked by the expert were merged and framed onto the image and also the coverage percentage of the markings was presented by a pie chart. The HIAE also provides easy navigation in the histopathology image database for rapid access to the image data. The atlas data prepared using the HIAE, overlaid on two colorectal histopathology images are shown in Figure 5.14(a), 5.14(b).

An important point to emphasize here is that the ground-truth data was collected and used for the purpose of evaluating the performance of the proposed histopathology slide labelling method. While such datasets are required for training conventional supervised classification methods, the quasi-supervised learning

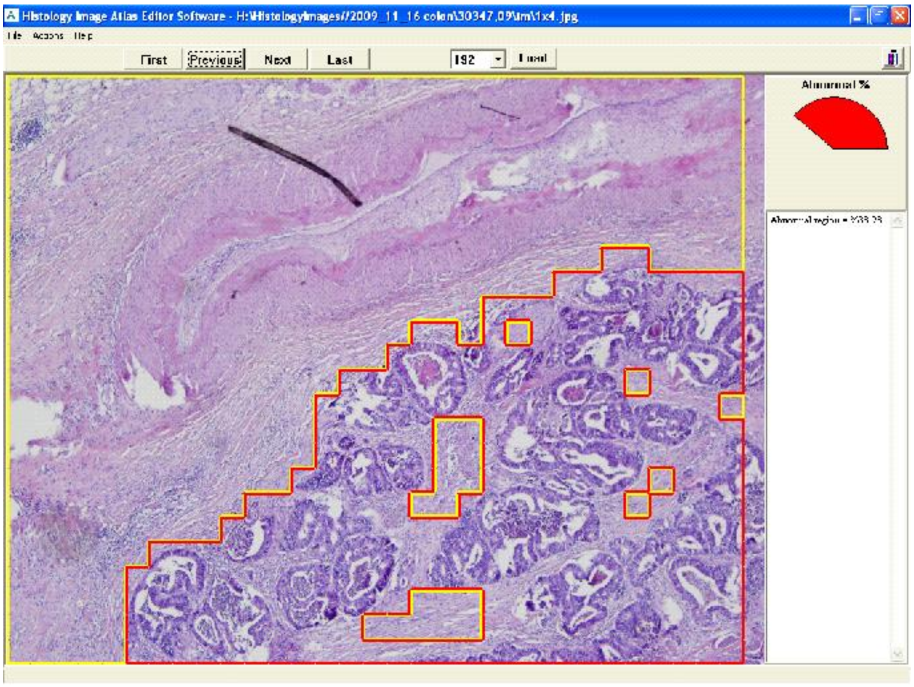

Histology Image Atlas Editor Software - H:\Histologyimages//2009_11_16 colon\30347,09\im\1x4.jpg
File Actions Help
First
Previous
Next
Last
192
Load
Abnormal %
Abnormal region = [illegible]

Figure 5.1. Histopathological Image Atlas Editor (HIAE) software tool screen snapshot.

paradigm adopted here is designed explicitly to remove the need for ground-truth training datasets. Hence, the ground-truth dataset was withhold from the quasi-supervised learning in the experiments.

### 5.2.2 Receiver Operating Characteristics Curve

In order to asses the separation of the NNCR and CRCa tissue regions, receiver operating characteristics (ROC) curves were constructed. An ROC curve is a graphical plot of the true positive rate versus the false positive rate achieved in a recognition experiment. The true positive rate, $P_{TP}$, denotes the probability of successful labelling of all ground-truth cancer features vectors. Similarly, the false positive rate, $P_{FP}$, denotes the probability of labelling NNCR features as cancer. In order to generate an ROC curve, we have computed $P_{FP}$ on the reference vector dataset and $P_{TP}$ on the unlabelled vector dataset, and plotted them for varying threshold $T$ ranging from 0.0 to 1.0.

We have evaluated the recognition performance of an ROC curve following two separate strategies:

1. The optimum recognition point on the ROC curve: The optimum threshold $T_{opt}$ is selected on the knee point of a continuous ROC curve where its slope equals to 1.0. In the ideal case, the ROC curve would be the unit step function and the optimum recognition point on this curve would be at the $(0.0, 1.0)$ point. This ideal point means that there is no false alarm with full true detection. After $T_{opt}$ is determined, the final labelling of the images was carried out using this threshold value and the corresponding $P_{FP}$ and $P_{TP}$ values were registered as the performance measures of that labelling experiment.

2. The area under the ROC curve: The area under the ROC curve is another performance measure in which a larger area means a better separation of the NNCR and CRCa tissue regions. In the ideal case described above, the area under the ROC curve would be equal to 1.0.

When comparing the results of the experiments, the one with the optimum recognition point $(P_{FP}, P_{TP})$ closest to ideal point $(0.0, 1.0)$ was identified as achieving a better identification performance. If two experiments had very close optimum recognition points, then the one with larger area under its ROC curve was accepted as offering more successful identification.

## 5.3 Texture Feature Vector Datasets

Throughout this study, various texture feature vector datasets were generated in order to evaluate the corresponding texture labelling and recognition performances. A texture feature dataset is differentiated with the parameters used in the feature vector extraction. Texture extraction configuration parameters used in this study are listed in groups as below:

1. The image plane,

   (a) Gray level image plane,

   (b) *Lab* color image planes.

2. Basic radius $r$ of the texture feature calculation geometry,

3. Pairwise pixel distance $d$,

   (a) Single $d$ value,

   (b) Multiple $d$ values.

4. Hierarchical feature computation,

   (a) Single neighbourhood ($H = 1$),

   (b) Multiple nested neighbourhoods ($H \geq 2$).

In the experiments, both the gray level and the color information obtained from the histopathological images were processed in parallel. In case of color image processing, the original images were transformed into the $Lab$ image planes using the well-known $RGB$ to $Lab$ color transformation using a white point of $(255, 255, 255)$ in the $RGB$ space, and each image plane was processed as a separate gray scale image (CIE 1986, Schwarz et al. 1987). Specifically, for an image region, texture feature vectors were calculated for each of $L$, $a$ and $b$ planes were then concatenated to produce a single texture feature vector. Therefore, a texture feature vector generated from an $Lab$ image had three times the dimension as the one obtained from a gray level image. In addition, uniform scalar quantization of the original 8-bit intensity levels was carried out into 16 quantized intensity levels on each image plane of interest before calculating the co-occurrence matrices to limit the number of possible image intensity pairs.

There have been several limitations in selecting a radius value $r$ for the feature calculation geometry. Firstly, the radius value was to match the discriminative local texture characteristics. Large radius values yield fewer feature vectors in total than the smaller radius values and sustain difficulties in defining the regions of texture transition. On the other hand, smaller radius values provide relatively higher separation in texture transition regions but make the labelling problem labour intensive due to a greater number of texture feature vectors. Last but not least, the choice of r had to take into account the smallest artefact observed in the Histopathological Image Library that is the cell nucleus ranging between $15-20$ microns in diameter. As a result, several different $r$ values; $32, 48, 64$ and 128 pixels were used in the texture extraction configuration. Matching values were taken into consideration for the pairwise pixel distance $d$.

Finally, the level of hierarchy $H$ was selected as either 1 or 2, limiting the feature vector computations to neighbourhoods of radii $r$ and $2r$. The case with $H = 1$ is also regarded as the "no hierarchy" case.

The resulting texture feature vector datasets of this study are listed in Table 5.1 with their respective texture feature extraction parameters and the corresponding vector dimensions:

Table 5.1. Texture feature vector datasets with corresponding texture extraction parameters.

| Texture Features | Image Planes | $r$ | $H$ | $d$ | Dimension |
|---|---|---|---|---|---|
| Dataset 1 | Gray | 64 | 2 | $\{1, 3, 5, 9, 13, 17, 21, 41, 51, 61\}$ | 296 |
| Dataset 2 | Lab | 64 | 2 | $\{1, 3, 5, 9, 13, 17, 21, 41, 51, 61\}$ | 888 |
| Dataset 3 | Lab | 64 | 1 | $\{1, 3, 5, 9, 13, 17, 21, 41, 51, 61\}$ | 444 |
| Dataset 4 | Lab | 48 | 2 | $\{1, 3, 7, 13\}$ | 384 |
| Dataset 5 | Lab | 64 | 2 | $\{1\}$ | 132 |
| Dataset 6 | Lab | 32 | 2 | $\{1, 3, 7, 13\}$ | 384 |
| Dataset 7 | Lab | 48 | 2 | $\{1\}$ | 132 |
| Dataset 8 | Lab | 48 | 1 | $\{1, 3, 7, 13\}$ | 192 |

## 5.4 Texture Labelling By The Quasi-supervised Learning

In our study, the QSL algorithm was operated on the texture feature vectors corresponding to the reference and mixed image groups and the subsequent labelling of the corresponding image regions were determined along with the optimal QSL reference set size $n_{opt}$.

The procedure provides the posterior probability estimates $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$ for all feature vectors in both groups computed for the optimal reference set size $n_{opt}$, with $p_0(\mathbf{x}) + p_1(\mathbf{x}) = 1$ for all feature vectors $\mathbf{x}$. In the labelling of a feature vector $\mathbf{x}$, a high $p_0(\mathbf{x})$ value means that the vector in question is more similar to those observed over NNCR tissues in the reference dataset. In turn, a low $p_0(\mathbf{x})$ value means that the vector in question is different from the NNCR feature vectors and by construction associated with CRCa.

In addition, during the computation of the posterior probabilities, the feature vectors obtained from the same image as the vector in consideration were left outside of the analysis in order to prevent biasing the analysis results due to the possible similarity between feature vectors obtained from the same histopathology slide.

### 5.4.1 Calculation Of The Optimal Reference Set Size

It should be highlighted again that the accuracy of the QSL algorithm is closely related to the reference set size $n$, which denotes the number of selected samples from each QSL group, to construct the random reference sets for the ensuing nearest neighbour classifier. In all of the texture labelling experiments, the first task was to determine the $n_{opt}$ value among all possible $n$'s that minimizes the cost functional $E(n)$ given in Equation 4.22. Using the $n_{opt}$ value, the posterior probabilities corresponding to each texture feature vector were calculated, as detailed in Section 4.2.1.

It was experimentally observed that the characteristics of the cost functional $E(n)$ with respect to $n$ was quite similar to the curve represented in Figure 5.2 in all of the texture labelling experiments operated on the Histopathological Image Library. For this $E(n)$ curve, an almost linear curve segment is observed for $n$ values beyond the global minimum point. This means that for these $n$ values, an accurate class separation was achieved but the penalty was being incurred primarily due to the increasing reference set size. In contrast, for $n$ values that are less than the optimum value, the class overlap was relatively higher, and thus, the first term of the cost functional was dominant.

It is very clear that it is not feasible to calculate all $E(n)$ values for all possible $n$ and determine the optimum value. On the other hand, the smooth nature of the $E(n)$ curve as illustrated above, allows determining $n_{opt}$ by using a Steepest-Descent algorithm. The algorithm starts from an initial estimate of $n$ and moves in the

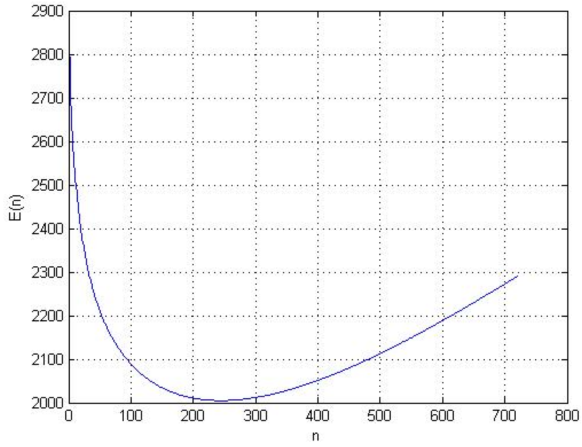


Figure 5.2. Common characteristics of cost functional $E(n)$ versus $n$.

negative direction of the gradient until no change occurs. The iterations starts with an initial estimate $n_0$ and the next reference set size value is calculated by

$$n_{k+1} = n_k - \lambda \frac{dE(n)}{dn} \text{ for k} = 0,1,2,\ldots \tag{5.1}$$

where $n_k$ and $n_{k+1}$ are the values in two successor iterations and $\lambda$ is a small positive coefficient that adjusts the convergence ratio. The algorithm is terminated when the absolute tangent of the $E(n)$ curve is below some predefined threshold, that was experimentally set to 0.05. Please also note that, due to the characteristics of the $E(n)$ curve the convergence would be faster if the initial $n$ values are selected relatively small since the curve is quite steep for small $n$ values.

The optimal reference set sizes determined for the QSL experiments performed on texture feature vectors Dataset 1 - 8 are given in the Table 5.2. For each texture feature vector dataset, the number of vectors in the reference group and the unlabelled mixed group, denoted by $\ell_0$ and $\ell_1$ respectively, are also shown in order to allow comparison in terms of the optimal reference set size $n_{opt}$.

### 5.4.2 Texture Labelling Performances

This section presents the texture labelling performances of the QSL labelling experiments for original texture feature vector datasets. The ROC curves obtained

Table 5.2. Optimum reference set size ($n_{opt}$) versus texture feature vector datasets. $\ell_0$ and $\ell_1$ represent the number of vectors in control group and mixed groups respectively.

| Texture Feature Dataset | $n_{opt}$ | $\ell_0$ | $\ell_1$ |
|---|---|---|---|
| Dataset 1 | 738 | 12192 | 9888 |
| Dataset 2 | 966 | 12192 | 9888 |
| Dataset 3 | 1401 | 17780 | 14420 |
| Dataset 4 | 2266 | 27432 | 22248 |
| Dataset 5 | 1025 | 12192 | 9888 |
| Dataset 6 | 8143 | 71120 | 57680 |
| Dataset 7 | 2115 | 27432 | 22248 |
| Dataset 8 | 2621 | 35560 | 28840 |

by the QSL algorithm operated on Datasets 1-6 are given in Figure 5.3. According to the texture labelling performance comparison rules described in Section 5.2 the best recognition performance was acquired with the texture vector Dataset 2. It was also observed that better optimum recognition was associated with larger ROC curve area.

Among the various datasets corresponding to different feature extraction parameters, the output labelling performance of the QSL method was the lowest when applied to Dataset 1 (see Table 5.1) constructed using the gray level image information. Conversely, the color texture feature information extracted from the *Lab* color space offered the best characterization of the NNCR and CRCa features. The QSL algorithm labelling performances calculated using these ROC curves were also given in Table 5.3.

In this table, $(P_{FP}, P_{TP})$ pairs are listed along with the area under ROC curve values. In terms of the ROC areas the worst texture labelling rate was obtained for Dataset 1 which is based on the gray level image information. The area under ROC values for vector datasets except Dataset 1 are very close to each other indicating very similar labelling performances. The optimum recognition points $(P_{FP}, P_{TP})$ also support this observation.

In general, $P_{TP}$ rates are quite satisfactory reaching the values up to 84%, while, the $P_{FP}$ rates are somewhat high. This phenomenon is referred to as the "malign tendency" of the algorithm, implying a general bias towards labelling NNCR

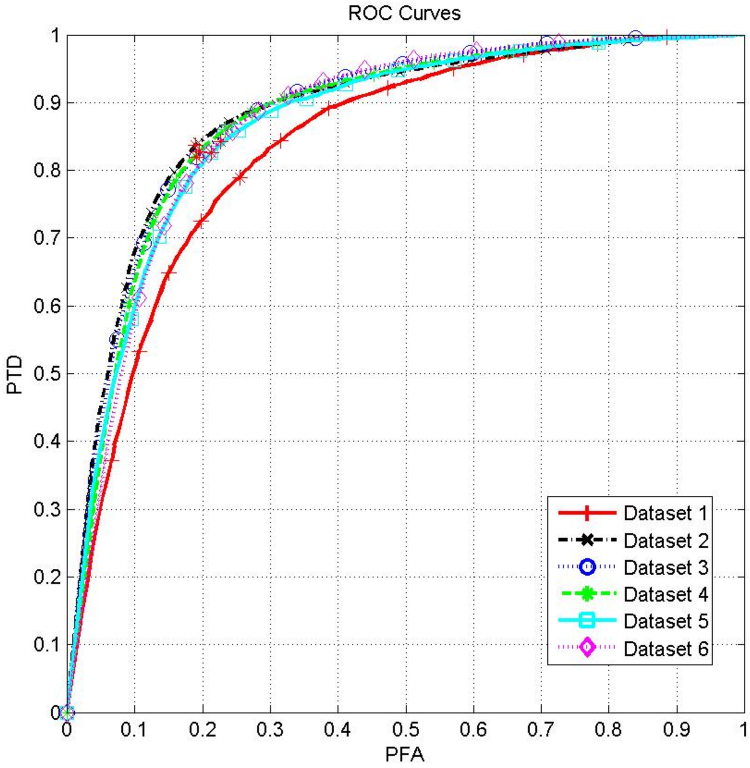


Figure 5.3. ROC curves (Dataset 1 - 6).

texture vectors as CRCa.

## 5.5 Texture Labelling Performances Versus Texture Extraction Parameters

This section presents the impact of the texture feature extraction parameter selection to the texture labelling performances. There are four different texture feature extraction parameters available as stated in Section 5.3. In this study, only one of the feature extraction parameters was changed while the other texture feature extraction parameters kept constant in order to observe its effects on the texture labelling performances.

In order to observe the effect of texture feature extraction parameter $H$ on the texture labelling performance, several QSL experiments were carried out on the texture feature datasets corresponding to $H = 1$ and 2, but for fixed $r$ and $d$. Texture labelling performances calculated for Dataset 2 versus Dataset 3 and Dataset 4

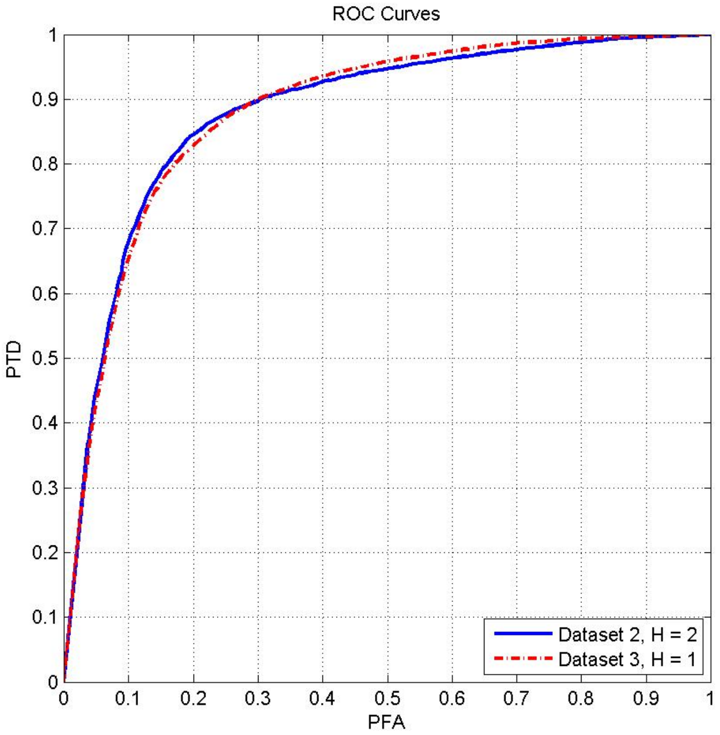
Table 5.3. Labelling performances of the QSL method for original texture feature vector datasets.

| Texture Features | $(P_{FP}, P_{TP})$ | ROC Area |
|---|---|---|
| Dataset 1 | (0.25, 0.79) | 0.84 |
| Dataset 2 | (0.19, 0.84) | 0.88 |
| Dataset 3 | (0.19, 0.82) | 0.88 |
| Dataset 4 | (0.20, 0.83) | 0.88 |
| Dataset 5 | (0.21, 0.83) | 0.87 |
| Dataset 6 | (0.23, 0.84) | 0.87 |
| Dataset 7 | (0.22, 0.84) | 0.87 |
| Dataset 8 | (0.20, 0.82) | 0.88 |

versus Dataset 8 were compared via several QSL experiments. The resulting ROC curves obtained for these experiments are given in Figure 5.4. By examining these ROC curve pairs and the corresponding QSL texture labelling performances given in Table 5.3, it can be concluded that varying values of parameter $H$ did not have any significant influence on the QSL texture labelling performance, with only a minor performance improvement of $P_{TD}$ values for $H = 2$.

In order to observe the effects of texture feature extraction parameter $r$ on the texture labelling performance, several QSL experiments were performed on the texture feature datasets corresponding to varying $r$ but fixed $H$ and $d$. To this end, texture labelling performances for Dataset 4 versus Dataset 6 and Dataset 5 versus Dataset 7 were compared. To this end, $r$ took the values of 48 and 32 for Datasets 4 - 6 and took the values of 64 and 48 for Datasets 5 - 7. Resulting ROC curves are given in Figure 5.5. By examining this ROC curve pairs and the corresponding QSL texture recognition performances given in Table 5.3, it can be concluded that varying $r$ across the set $\{32, 48, 64\}$ did not have any significant influence on the QSL texture labelling performance.

Similarly, in order to observe the effects of texture feature extraction parameter $d$ on the texture recognition performance, several QSL experiments were carried out on the texture feature datasets corresponding to varying $d$ but fixed $H$ and $r$. Now, we describe two sample labelling experiment couples; in the first couple, the experiments were carried out on features vectors Dataset 2 and Dataset 5. In the first experiment of this couple, texture feature vectors were calculated on Dataset



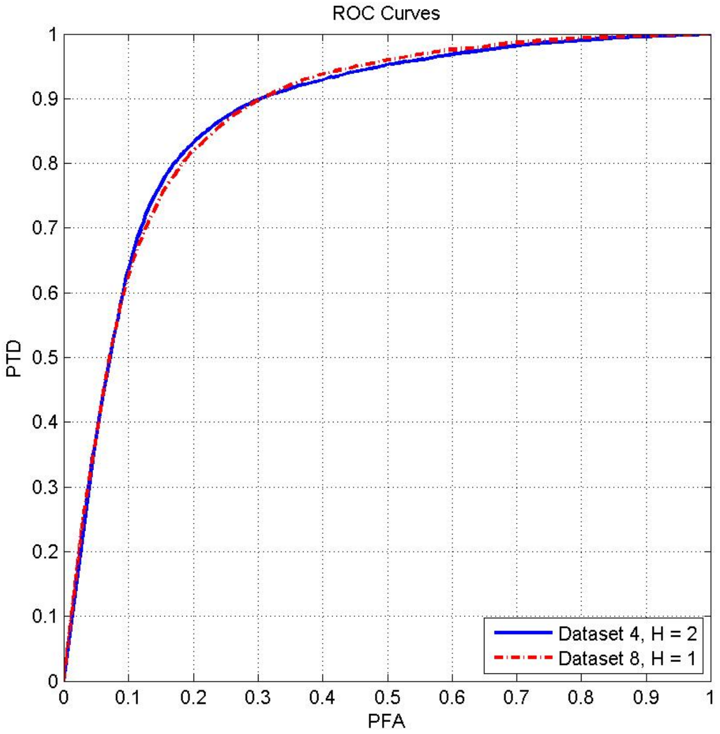
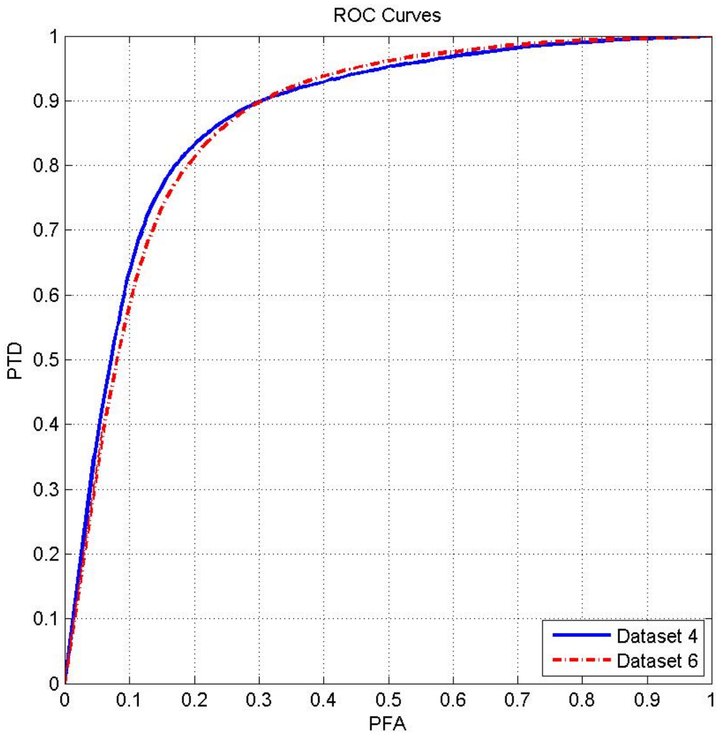


Figure 5.4. Comparisons of ROC curve pairs for variable texture feature extraction parameter $H$.



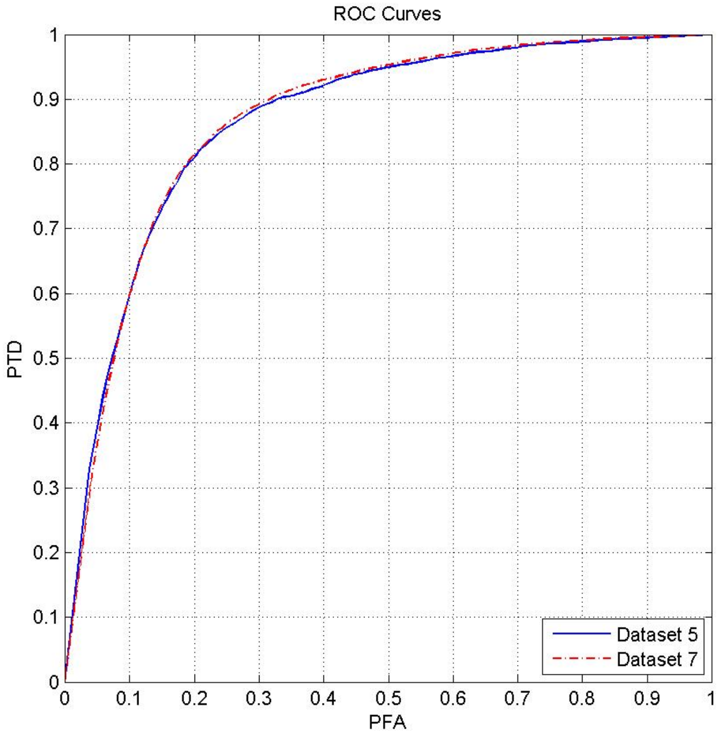


Figure 5.5. Comparisons of ROC curve pairs for variable texture feature extraction parameter $r$. (For corresponding feature extraction parameters see Table 5.1)

2 for $d \in \{1, 3, 5, 9, 13, 17, 21, 41, 51, 61\}$. In the second experiment, texture feature vectors were calculated on Dataset 5 for $d = 1$. In these two experiments, $H = 2$ and $r = 64$ values kept the same.

For the next experiment couple, the two experiments were carried out on Dataset 4 and Dataset 7 separately. In the first experiment of this couple, texture feature vectors on Dataset 4 were calculated for $d \in \{1, 3, 7, 13\}$. In the other experiment, texture feature vectors on Dataset 7 were calculated for $d = 1$. In both experiments, $H = 2$ and $r = 48$ values kept unchanged.

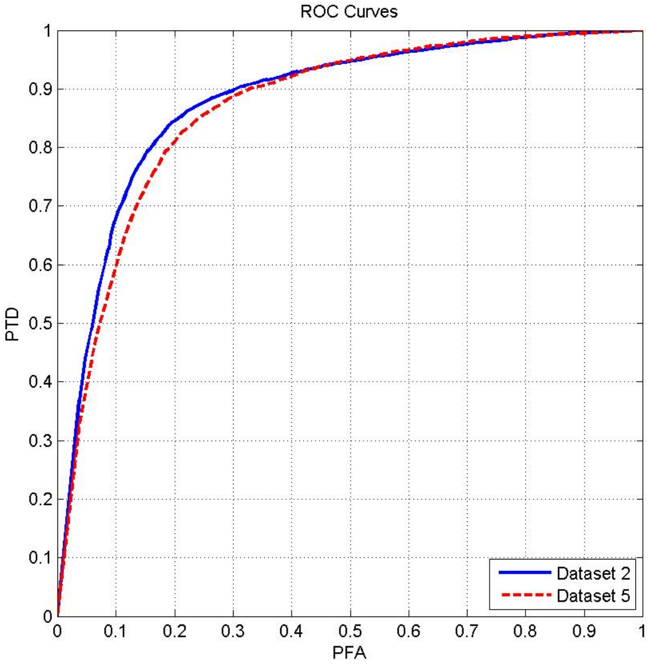
The resulting ROC curves for two experiment couples are shown in Figure 5.6. Examining the ROC curve pair given in Figure 5.6(a), it can easily be seen that the QSL labelling performance calculated for Dataset 2 is much better than of Dataset 5. In addition, the ROC curve pair shown in Figure 5.6(b) states that Dataset 4 had higher labelling performance compared to Dataset 7 but the improvement was relatively minor compared to the other experiment pair.

Dataset 2 experiment resulted with a larger area under the ROC curve compared to Dataset 5 (0.88 versus 0.87) and had successful optimum recognition point (0.18, 0.84) versus (0.21, 0.83); providing lower $P_{FA}$ coupled with a higher $P_{TD}$. Dataset 4 experiment also resulted with a larger area under ROC compared to Dataset 7 (0.88 versus 0.87) and a lower $P_{FA}$ value at the optimum labelling threshold. For the performance measures please refer to Table 5.3.

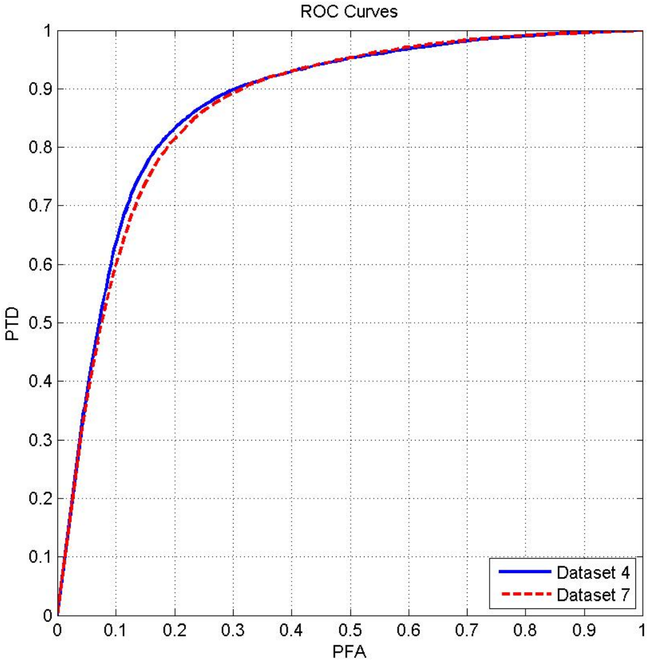
As a result of the comparisons detailed above, it can be concluded that QSL labelling performances for texture feature vector datasets calculated combining multiple inter-pixel distances were better than that calculated for $d = 1$ only. This means that including one or more inter-pixel distance values in addition to $d = 1$ increased the separation among healthy and cancer texture feature vectors.

## 5.6 Comparative Labelling Case: A Support Vector Machine Recognition

It is very important compare the QSL method against other off-the-shelf vector classifiers on the same texture feature vector datasets in order to comment about the labelling performance of QSL. In order to obtain an independent evaluation of



(a)



(b)

Figure 5.6. Comparisons of two ROC curve pairs corresponding to varying texture feature extraction parameter $d$. (For corresponding feature extraction parameters see Table 5.1)

the labelling performance, we have used a Support Vector Machine (SVM) classifier trained on the ground-truth label data that was withheld from the quasi-supervised labelling strategy (Cortes and Vapnik 1995, Vapnik 1998, Burges 1998).

In order to perform the SVM classification experiments, we constructed the NNCR and CRCa vector groups using ground-truth atlas information. Group vectors took labels $+1$ and $-1$, depending on whether they belong to the NNCR group or not. An SVM classifier training vector set was then constructed by randomly selecting 90% of these vectors along with a control vector set constructed by the remaining 10%. The training feature vector set was used to obtain the classifier model that was then operated on the control feature vector set. The SVM classifier output obtained for the vectors in the control set was compared with the ground truth labels in the Histopathological Image Atlas to calculate the resulting classification performance.

Since a relatively smaller number of texture vectors was used in the control set for an individual SVM experiment, the performance measures obtained were not expected to represent the actual classifier performance well. To address this issue, multiple independent SVM classification experiments (40 in our study) were performed for a specific texture vector dataset and the resulting classification performances were used to determine the corresponding statistics via their means and their variances. The mean and the variance statistics determined the performance of the SVM classifier on a specific texture vector dataset.

In the SVM classification experiments, we have used a third party implementation, $SVM^{light}$ (http://svmlight.joachims.org). For the background on $SVM^{light}$ please see Section 4.3.4.

In the SVM classifier construction, we have used a Gaussian Radial Basis function kernel (see Equation 4.49) where the scale parameter $\sigma$ was determined by minimizing the number of Support Vectors in the training phase via a line search. In order to take into account the non-separable cases, the Lagrange multipliers of the quadratic optimization were bound from above by 100.0 during training, producing a soft-margin classification. The recognition performance of the samples in the control dataset was carried out by thresholding the classifier underlying function

$$h(x) = \sum_i y_i \beta_i K(x, x_j) + b \tag{5.2}$$

by a threshold $T \in (-\infty, +\infty)$, $y_i$ being $+1$ or $-1$ based on whether $x_i$ belongs to the NNCR or CRCa groups respectively, with $\beta_i$ and $b$ obtained during the training of the classifier. The control dataset samples for which $h(x) \geq T$ were then recognized as NNCR.

The QSL algorithm labelling performances and the SVM classifier performances for texture feature Datasets 1-6 are presented in Table 5.4. Note that, while labelling performance measure values are given directly for the QSL experiments, the classification performance statistics are given for the SVM experiments as explained. It can be observed that the SVM classifier probability of false positive ($P_{FP}$) values are distributed around relatively lower values compared to that of QSL. while the probability of true positive ($P_{TP}$) values follow each other very closely. Note, however, that as the SVM classifier was trained on the atlas data, it represents an upper bound to the QSL labelling performance in the QSL application, since there is no ground-truth data to be used in learning. The results in Table 5.4 indicate that the QSL strategy attains a labelling performance that is close to this upper bound in terms of true positive rate without the benefit of a ground-truth learning dataset.

Table 5.4. Performances of the QSL method with the SVM classifier performance statistics.

| | QSL | SVM | |
|---|---|---|---|
| Texture Features | $(P_{FP}, P_{TP})$ | $P_{FP}(\mu, \sigma^2)$ | $P_{TP}(\mu, \sigma^2)$ |
| Dataset 1 | (0.25, 0.79) | (0.15, 0.01) | (0.87, 0.01) |
| Dataset 2 | (0.19, 0.84) | (0.03, 0.00) | (0.96, 0.00) |
| Dataset 3 | (0.19, 0.82) | (0.03, 0.00) | (0.94, 0.00) |
| Dataset 4 | (0.20, 0.83) | (0.04, 0.02) | (0.86, 0.26) |
| Dataset 5 | (0.21, 0.83) | (0.04, 0.03) | (0.84, 0.40) |
| Dataset 6 | (0.23, 0.84) | (0.02, 0.00) | (0.96, 0.00) |

## 5.7 Vector Dimensionality Reduction

As explained in detail in Section 4.4, vector dimensionality reduction is a mathematical transformation to represent a vector dataset in a relatively lower dimensional space. One of the advantages of a reduced dimensionality is alleviating the Curse of Dimensionality (see Section 4.4.1), leading to better classification performances and reduced computation times.

This section presents the dimensionality reduction strategies performed on the texture feature datasets and the experimental results obtained using the reduced datasets. The main purpose of this work was to evaluate all possible dimensionality reduction strategies along with their adaptation parameters. However, because of the high number of texture feature vectors and the high texture feature vector dimensions, it was not possible to evaluate all the dimensionality reduction procedures explained in Section 4.4. In addition, it was also not possible to evaluate all the variations for a specific dimensionality reduction method, such as, selecting various subsets for feature subset selection procedure and selecting various target dimensionalities for PCA procedure. Despite these limitations, a comprehensive study was conducted by evaluating the following dimensionality reduction methods;

1. Individual Feature Selection,

2. Principal Component Analysis,

3. Isomap,

4. Independent Component Analysis,

presented in detail in the following sections.

### 5.7.1 Individual Feature Selection

Individual feature selection is a methodology where a subset of features is selected individually from the complete feature dataset according to a defined selection criterion (see Section 4.4.3). In our study, the purpose of the subset selection was to have higher QSL classification performances using the reduced feature vector datasets. In an ideal feature selection case, exhaustive selection would be performed, in other words all possible feature combinations would be selected and the QSL performances would be evaluated for each of them. However, this strategy is generally not applicable since the number of distinct texture feature combinations and hence the total processing time is typically very high. As a result, a much more limited approach was conducted in this study, only a finite number of combinations of texture features were selected and the QSL method was carried out on them. The main idea here was to detect texture features that provided a higher separation between the reference and the mixed groups individually and to use several combinations of these components instead of the original set of texture features.

In our study, the first step was to calculate the level of separation for each individual component of an original texture feature vector dataset. A measure of separation $s_i$ for the $i$'th texture feature vector component was defined as;

$$s_i = \frac{|\mu_{iR} - \mu_{iM}|}{\sigma_{iR}^2 + \sigma_{iM}^2} \tag{5.3}$$

where $\mu_{iR}$ and $\mu_{iM}$ denote the mean values and $\sigma_{iR}$ and $\sigma_{iM}$ the standard deviations of the $i$'th feature vector component for the reference and the mixed vector groups respectively. The measure of separation implies that the statistical distribution of the two observations are well separated if the mean distance of these observations is relatively larger, together with a smaller observation-specific variances.

Suppose that, between two different texture features $t_1$ and $t_2$ ,$t_1$ has a better separated distribution than $t_2$ for healthy and cancer labels; by means of its larger distance between group means and by means of smaller group variances. Considering the QSL experiment groups, which are healthy and mixed, the mixed group vector component distribution would be equal to the sum of partial distributions of observations from both healthy and cancer feature vectors. On the other hand, the reference group would have only a healthy vector component distribution for components $t_1$ and $t_2$. This implies that

$$|\mu_{t_1R} - \mu_{t_1M}| > |\mu_{t_2R} - \mu_{t_2M}| \tag{5.4}$$

$$\sigma^2_{t_1R} < \sigma^2_{t_2R} \tag{5.5}$$

$$\sigma^2_{t_1M} < \sigma^2_{t_2M} \tag{5.6}$$

resulting with :

$$s_{t_1} > s_{t_2} \tag{5.7}$$

In order to examine the separation of texture features in a feature vector dataset, the criterion in Equation 5.3 was calculated for each individual texture feature and these features are sorted in a descending order. By this sorted list, texture feature characteristics and the corresponding texture feature extraction parameters could be compared with each other and the parameters that yielded higher separations could be determined. The texture feature components from Datasets 2, 5 and 1 (see Table 5.1) that have best and worst ten measures of separation are listed in Table 5.5 - 5.10.

By examining all of the lists of the best ten texture vector components, it can be summarized that:

1. All of the best texture vector components are second order texture feature characteristics. There is no first order texture feature characteristic in the list.
2. For the texture features extracted using the *Lab* color planes, texture feature vector components calculated using color plane *a* were better.
3. The majority of these texture features were calculated for $d = 1$.

The texture feature characteristics that exhibited higher measures of separation were;

1. Inverse difference moment,
2. Difference variance,
3. Difference entropy.

Table 5.5. The texture feature characteristics (from Dataset 2) that have the highest ten measure of separation and corresponding texture extraction parameters.

| Image Plane | $H$ | $r$ | $d$ | FOS | SOS | Texture Features |
|---|---|---|---|---|---|---|
| a | 2 | 128 | 1 | | √ | Inverse Diference Moment |
| a | 2 | 128 | 1 | | √ | Difference Variance |
| a | 2 | 128 | 1 | | √ | Difference Entropy |
| b | 2 | 128 | 1 | | √ | Difference Entropy |
| L | 2 | 128 | 1 | | √ | Inverse Difference Moment |
| b | 2 | 128 | 1 | | √ | Difference Variance |
| b | 2 | 128 | 3 | | √ | Difference Entropy |
| b | 2 | 128 | 1 | | √ | Inverse Difference Moment |
| b | 2 | 128 | 1 | | √ | Entropy |
| b | 2 | 128 | 3 | | √ | Difference Variance |

Table 5.6. The texture feature characteristics (from Dataset 2) that have the lowest ten measure of separation and corresponding texture extraction parameters.

| Image Plane | $H$ | $r$ | $d$ | FOS | SOS | Texture Features |
|---|---|---|---|---|---|---|
| a | 1 | 64 | | √ | | CPVMVH |
| L | 1 | 64 | | √ | | Kurtosis |
| L | 1 | 64 | 61 | | √ | IMC2 |
| a | 1 | 64 | 41 | | √ | IMC2 |
| L | 1 | 64 | 51 | | √ | IMC2 |
| L | 2 | 128 | | √ | | CPVMVH |
| a | 1 | 64 | 61 | | √ | IMC1 |
| b | 1 | 64 | 41 | | √ | IMC1 |
| L | 1 | 64 | | √ | | CPVMVH |
| a | 1 | 64 | 51 | | √ | IMC1 |

Table 5.7. The texture feature characteristics (from Dataset 5) that have the highest first ten measure of separation and corresponding texture extraction parameters.

| Image Plane | $H$ | $r$ | $d$ | FOS | SOS | Texture Features |
|---|---|---|---|---|---|---|
| a | 2 | 128 | 1 | | √ | Inverse Diference Moment |
| a | 2 | 128 | 1 | | √ | Difference Variance |
| a | 2 | 128 | 1 | | √ | Difference Entropy |
| b | 2 | 128 | 1 | | √ | Difference Entropy |
| L | 2 | 128 | 1 | | √ | Inverse Difference Moment |
| b | 2 | 128 | 1 | | √ | Difference Variance |
| b | 2 | 128 | 1 | | √ | Inverse Difference Moment |
| b | 2 | 128 | 1 | | √ | Entropy |
| L | 2 | 128 | 1 | | √ | IMC1 |
| a | 1 | 64 | 1 | | √ | Difference Entropy |

Table 5.8. The texture feature characteristics (from Dataset 5) that have the lowest last ten measure of separation and corresponding texture extraction parameters.

| Image Plane | $H$ | $r$ | $d$ | FOS | SOS | Texture Features |
|---|---|---|---|---|---|---|
| a | 2 | 128 | | √ | | CPVMVH |
| b | 1 | 64 | | √ | | Kurtosis |
| b | 2 | 128 | | √ | | CPVMVH |
| L | 2 | 128 | | √ | | Kurtosis |
| b | 2 | 128 | | √ | | Kurtosis |
| b | 1 | 64 | | √ | | CPVMVH |
| a | 1 | 64 | | √ | | CPVMVH |
| L | 1 | 128 | | √ | | Kurtosis |
| L | 2 | 128 | | √ | | CPVMVH |
| L | 1 | 64 | | √ | | CPVMVH |

Table 5.9. The texture feature characteristics (Dataset 1) that have the highest first ten measure of separation and corresponding texture extraction parameters.

| Image Plane | $H$ | $r$ | $d$ | FOS | SOS | Texture Features |
|---|---|---|---|---|---|---|
| Gray Level | 2 | 128 | 1 | | $\surd$ | Inverse Diference Moment |
| Gray Level | 2 | 128 | 1 | | $\surd$ | Difference Entropy |
| Gray Level | 2 | 128 | 1 | | $\surd$ | Difference Variance |
| Gray Level | 2 | 128 | 1 | | $\surd$ | IMC1 |
| Gray Level | 1 | 64 | 1 | | $\surd$ | Difference Entropy |
| Gray Level | 2 | 128 | 1 | | $\surd$ | Entropy |
| Gray Level | 1 | 64 | 1 | | $\surd$ | Inverse Difference Moment |
| Gray Level | 2 | 128 | 3 | | $\surd$ | Difference Entropy |
| Gray Level | 1 | 64 | 1 | | $\surd$ | Difference Variance |
| Gray Level | 2 | 128 | 3 | | $\surd$ | Inverse Difference Moment |

Table 5.10. The texture feature characteristics (Dataset 1) that have the lowest last ten measure of separation and corresponding texture extraction parameters.

| Image Plane | $H$ | $r$ | $d$ | FOS | SOS | Texture Features |
|---|---|---|---|---|---|---|
| Gray Level | 1 | 64 | 61 | | $\surd$ | Angular Second Moment |
| Gray Level | 1 | 64 | 61 | | $\surd$ | IMC1 |
| Gray Level | 2 | 128 | | $\surd$ | | Kurtosis |
| Gray Level | 1 | 64 | 51 | | $\surd$ | IMC1 |
| Gray Level | 1 | 64 | | $\surd$ | | Kurtosis |
| Gray Level | 1 | 64 | 41 | | $\surd$ | IMC2 |
| Gray Level | 1 | 64 | 51 | | $\surd$ | IMC2 |
| Gray Level | 2 | 128 | | $\surd$ | | CPVMVH |
| Gray Level | 1 | 64 | | $\surd$ | | CPVMVH |
| Gray Level | 1 | 64 | 61 | | $\surd$ | IMC2 |

Similarly, by examining the lists of the worst ten texture vector components, it can be summarized that:

1. Most of the worst texture vector components are first order texture features characteristics.
2. All of these texture feature vector components were extracted using relatively larger values of $d$, such as $d = \{41, 51, 61\}$.

The texture feature characteristics which have poor measure of separation were;

1. Corresponding Pixel Value for Maximum Value of Histogram (CPVMVH),
2. Kurtosis,
3. Information Measures of Correlation 1 (IMC1),
4. Information Measures of Correlation 2 (IMC2).

Since the second order feature characteristics have higher measures of separation, QSL experiments were repeated using only the second order feature characteristics. The resulting labelling performances were compared to each other as follows:

1. A baseline experiment was carried out on the original complete texture feature vector dataset.
2. A second experiment was carried out on only the first order features.
3. A third experiment was carried out on only the second order features.

The resulting ROC curves are given in Figure 5.7. The ROC curve obtained from the second order features is above the one obtained using the first order features, supporting the observation on the superior separability of the two regions using ordered measures of separation. On the other hand, the ROC curve obtained from the second order feature characteristics alone is surpassed by the original ROC curve. This means that better labelling performance was reached by concatenating both the first order and the second order texture feature vector feature subsets as opposed to using them individually. In all other texture feature vector datasets the same phenomenon was observed, hence, the ROC curves obtained using datasets other than Dataset 5 are not represented.

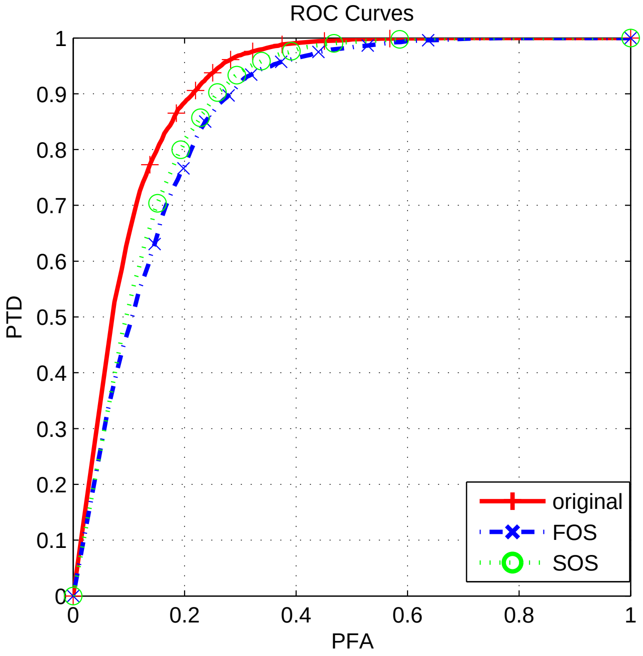


Figure 5.7. ROC curves for only the first order, only the second order and all characteristics together (Dataset 5).

In order to put these observations into a perspective, let us assume that a texture feature vector dataset has a well separated distribution between the reference and the mixed groups. This means that, this texture feature vector dataset can provide high QSL labelling performance. Another individual texture feature may also have a high separation performance and may potentially improve the overall QSL performance when concatenated to the first feature. However, the second individual texture feature may also carry redundant information to the first one, in which case the overall QSL labelling performance may stay more-or-less the same or may even decrease due to the increase in vector dimensionality due to the Curse of Dimensionality (see Section 4.4.1). Alternatively, the second texture feature may have lower separation performance and cause a drop in the overall recognition performance when it is concatenated to the original vector dataset.

In order to test this behaviour, several QSL experiments were performed and the results were compared to each other. These experiments were composed of a baseline experiment carried out on an original texture feature vector dataset followed by several related experiments in which the QSL algorithm was operated on different texture the feature subsets.

In a sample baseline experiment, which that was carried out on Dataset 5, the number of texture features was 132. After measure of separation values were calculated for the features, the first highest 10, 45, 80 and 120 components were selected in order to obtain secondary texture feature vector subsets with reduced dimensions. Subsequently, the QSL method was operated on each vector subset and the labelling results were compared to each other. The corresponding ROC curves are given in Figure 5.8 where the first 10 components with the largest separation scores achieved the worst classification performance of all. As the dimensions of the vector component subset increased, the ROC curves moved towards the ROC curve of the baseline experiment but never exceeded it. The same behaviour was also experimentally observed for several other vector datasets whose results are not shown here due to their similarity to this experiment.

### 5.7.2 Dimensionality Reduction By The Principal Component Analysis

Several texture classification experiments were performed to test whether the PCA procedure improved the labelling performance of the QSL method. As an example, QSL method was operated on several reduced texture feature vector datasets were generated using the PCA features. Next, the ROC curves were compared with each other.

The reduced vector datasets of dimensionality $1, 2, 4, 8, 16, 32$ and $64$ were calculated by PCA algorithm. Resultant ROC curves obtained for the baseline and the reduced dimensionality experiments are provided in Figure 5.9.

The results indicate that PCA dimensionality reduction method did not improve the QSL labelling performance. The ROC curve of the baseline experiment exceeded all other ROC curves obtained using PCA features. Another observa-

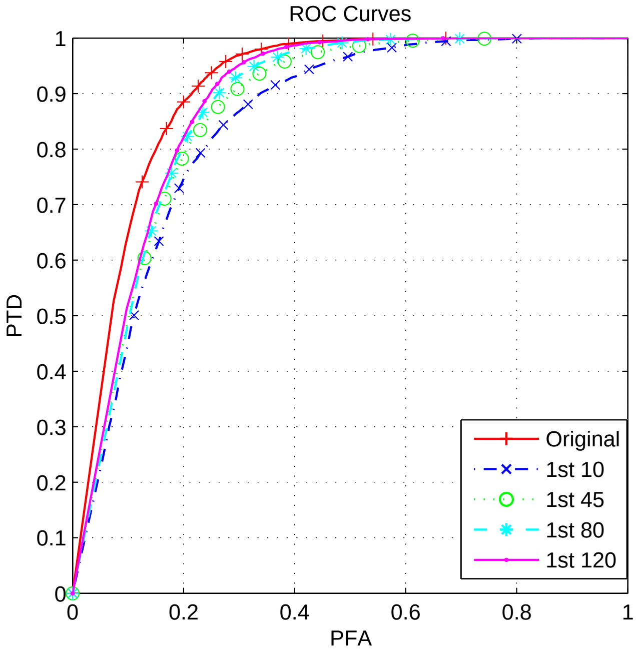


Figure 5.8. ROC curves for the original and subset of texture features obtained by selecting the list of the ones corresponding to higher measure of separation (Dataset 5).

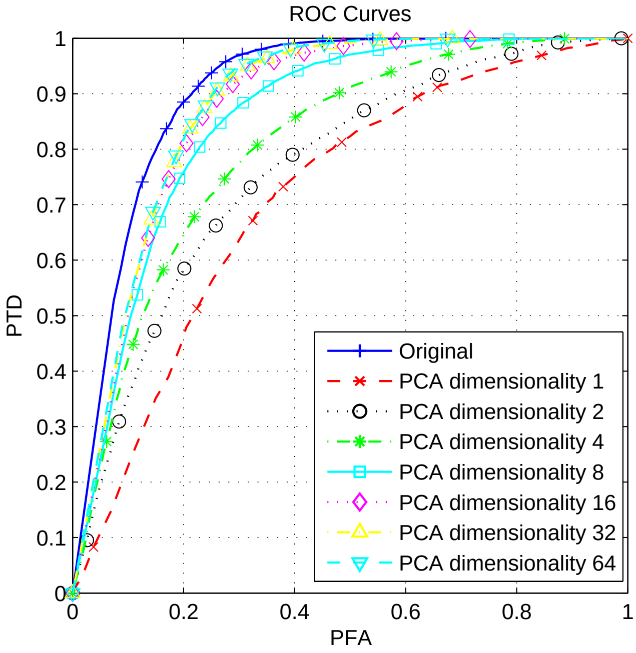


Figure 5.9. ROC curves for the original and PCA applied texture feature vectors (Dataset 5).

tion was that the minimum PCA dimensionality, which was 1 in this case, resulted with the worst labelling performance. As the PCA dimensionality increased, the ROC curves moved closer to the ROC of the baseline experiment but never reaching it. The same behaviour was also experimentally observed for other feature vector datasets.

### 5.7.3 Dimensionality Reduction By The Isomap Algorithm

Several texture labelling experiments were performed in order to assess if the Isomap dimensionality reduction approach improved the labelling performance of the QSL method. The methodology was, as usual, to operate the QSL method on the reduced texture feature vector datasets via the Isomap algorithm and to compare the resulting ROC curves with the one obtained on the original data previously.

The reduced vector datasets of dimensionality $3, 5, 8, 12$ and $16$ were calculated by the coordinates generated for the original data at the respective dimensions by the Isomap algorithm. The resultant ROC curves are shown in Figure 5.10 along with the ROC curve of the baseline experiment.

These results show that the Isomap dimensionality reduction method did not improve the QSL labelling performance either. The ROC curve of the baseline experiment bounded all other ROC curves from above. Furthermore, the minimum Isomap dimensionality, which was 3, produced the worst labelling performance. As the Isomap dimensionality increased, the ROC curves moved towards the ROC of the baseline experiment. Again, the same behaviour was also experimentally observed for several other texture feature vector datasets.

### 5.7.4 Visualisation Of The Texture Feature Vectors

While it would be very useful to investigate texture vectors visually in a vector space it is quite difficult due to high vector dimensionality. Visual investigation may give us ideas about the structure of the texture data and about the degree of the separation among specific vector labels. In the problem studied here, we used

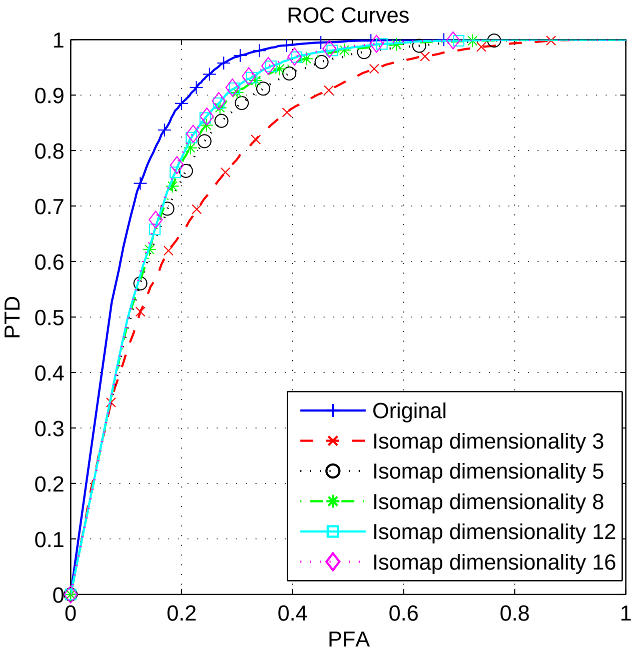


Figure 5.10. ROC curves for the original and Isomap applied texture feature vectors (Dataset 5).

the Isomap dimensionality reduction method to visually observe the distribution of texture feature vectors due to its topology preserving behaviour when a high dimensionality original texture feature vector dataset is reduced to a lower dimensionality (see Section 4.4.5). In particular, the embedding of the feature vector data to a three dimensional space allowed us to observe the feature vector distribution even though it still required some degree of manual graphical operations to highlight the local separation of the feature vectors and to observe the vectors in occluded areas. As a result simpler views were generated by Isomap dimensionality 2 that embeds the high dimensional texture feature vector space into thee $x - y$ plane.

Firstly, the texture feature vectors generated from the mixed group images were investigated in a two dimensional embedding. Since the mixed group images had CRCa tissues marked by an expert pathologist using the HIAE tool, it was possible to plot the ground-truth cancer vectors and the NNCR vectors differently. A plot of CRCa texture vectors and the rest for the mixed group is provided in Figure 5.11. The figure shows a partial overlap between the texture feature vectors of adenocarinoma and the rest in spite of a clear differentiation between them in the manual labelling.

Secondly, all texture vectors from both groups, namely, the reference and the mixed groups were plotted (see Figure 5.12). The figure shows the ground-truth information, the NNCR vectors from both the reference and the mixed groups and the CRCa vectors of the mixed group. Naturally, the partial overlap between NNCR and cancerous vectors are still observed.

### 5.7.5 Dimensionality Reduction By The Independent Component Analysis

In these experimentals, the original texture feature vector datasets were converted to reduced dimension feature vector datasets using the FastICA method (see Section 4.4.8). The reduced vector datasets were then fed to the QSL algorithm and the results were compared to those of the baseline experiments.

The baseline and ICA features vector dimensionality values together with labelling performances evaluated in experiments on Datasets 1 through 6 are presented in Table 5.11. The results indicate that in all of the ICA operated experiments, except the one on the Datasets 6, the ICA method improved the baseline labelling performance; increased the areas under the ROC curves and the optimum recognition points on the ROC curves moved closer towards the ideal point.

Table 5.11. Labelling performances of the QSL method for original and the ICA applied texture feature vector datasets.

| | Original | | | ICA | | |
|---|---|---|---|---|---|---|
| Texture Features | Dim. | $(P_{FP}, P_{TP})$ | ROC Area | Dim. | $(P_{FP}, P_{TP})$ | ROC Area |
| Dataset 1 | 296 | (0.25, 0.79) | 0.84 | 279 | (0.21, 0.84) | 0.89 |
| Dataset 2 | 888 | (0.19, 0.84) | 0.88 | 805 | (0.22, 0.88) | 0.90 |
| Dataset 3 | 444 | (0.19, 0.82) | 0.88 | 402 | (0.21, 0.86) | 0.89 |
| Dataset 4 | 384 | (0.20, 0.83) | 0.88 | 341 | (0.18, 0.86) | 0.91 |
| Dataset 5 | 132 | (0.21, 0.83) | 0.87 | 119 | (0.19, 0.88) | 0.91 |
| Dataset 6 | 384 | (0.23, 0.84) | 0.87 | 362 | (0.32, 0.98) | 0.86 |

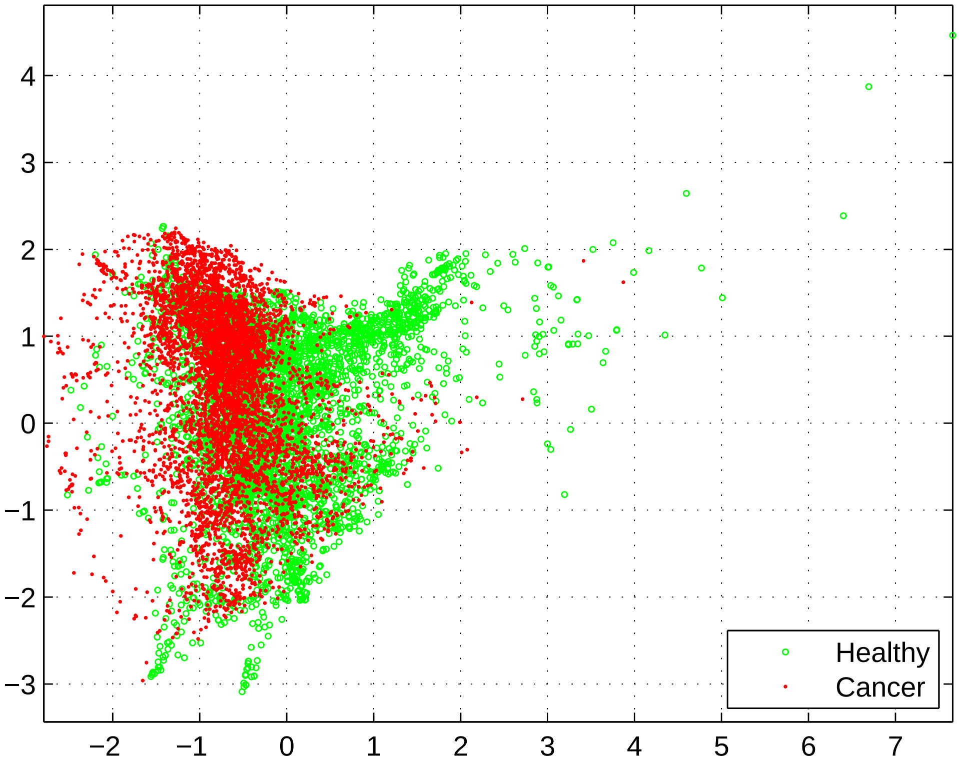


Figure 5.11. Dimensionally reduced texture feature vectors of the mixed group obtained by the Isomap algorithm. Red points represent CRCa, green points represent NNCR.

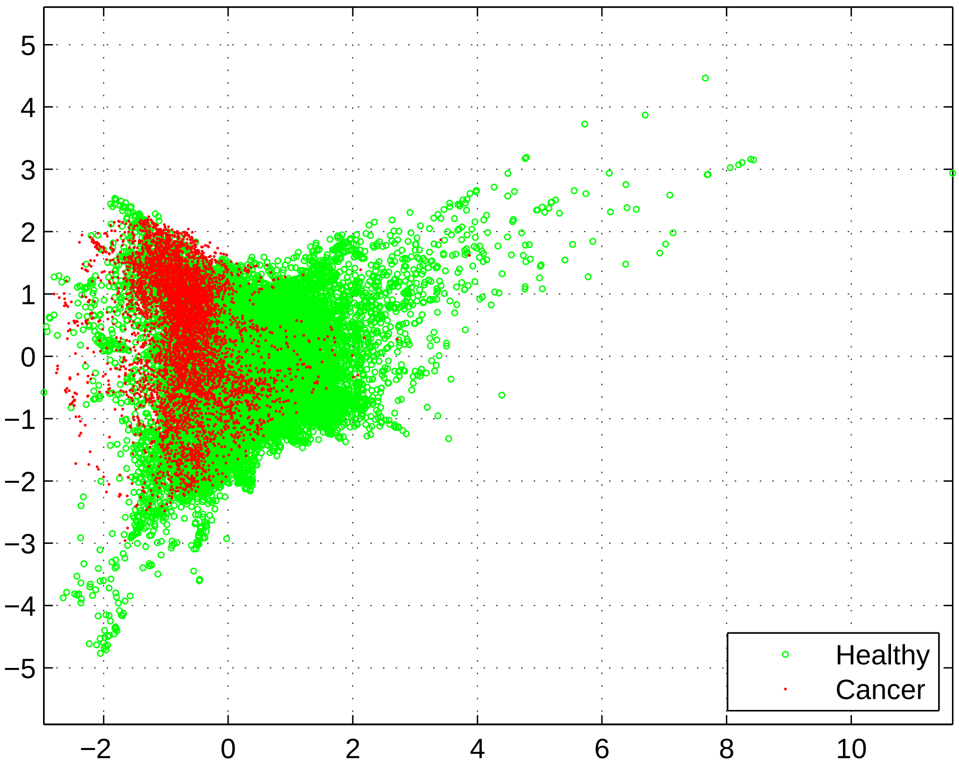
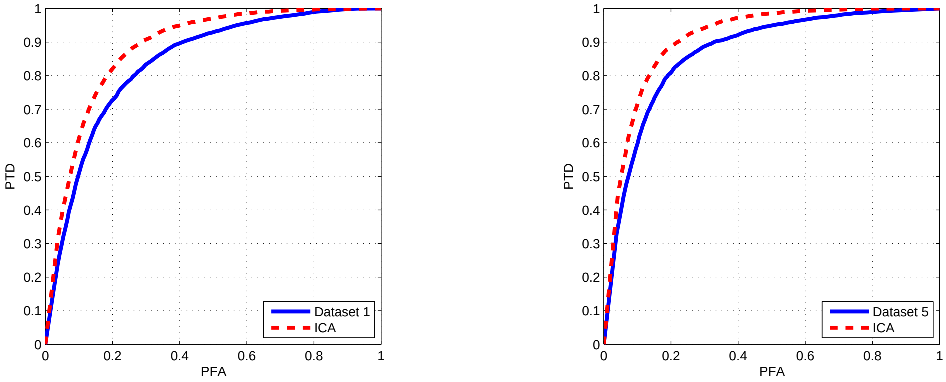


Figure 5.12. Dimensionally reduced texture feature vectors of the reference and the mixed groups generated by the Isomap algorithm. Red points represent CRCa, green points represent NNCR.

The ROC curves obtained from the baseline and the ICA experiments are given in Figure 5.13. In contrast to all previously considered dimensionality reduction techniques, the ROC curves obtained from the ICA vector datasets resulted with higher labelling performances compared to the baseline ROC curve. Similar improvements were also observed on the other texture feature vector datasets.



(a)

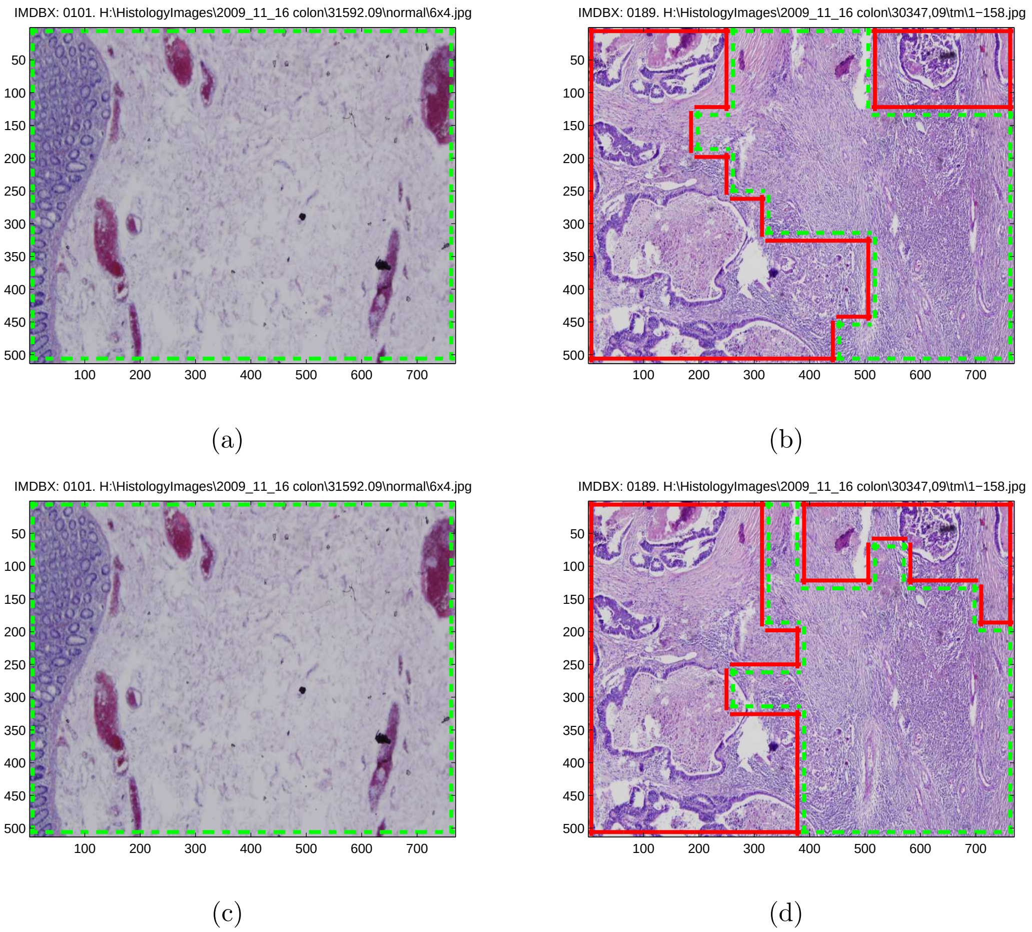
(b)

Figure 5.13. ROC curve pairs for the baseline and the ICA applied vector datasets. The baseline and the reduced vector dimensions for (a) Dataset1 are 296 and 279 (b) Dataset5 are 132 and 119.

A sample labelling of the QSL method following the ICA procedure overlaid on the histopathology images is presented in Figure 5.14(c), 5.14(d). In these figures, individual square regions that were identically labelled were merged and framed as an overlay onto the underlying histopathology images. The ground truth atlas data corresponding to these images are also presented in Figure 5.14(a), 5.14(b) and can be compared with the automated labelling results

## 5.8 Synthetic Validation Of The Quasi-supervised Learning Algorithm

To assess the QSL texture labelling performance on well separated healthy and cancer vector datasets, several experiments were carried out on the synthetic texture feature vectors. The synthetic feature vector components for NNCR and


IMDBX: 0101. H:\HistologyImages\2009_11_16 colon\31592.09\normal\6x4.jpg
IMDBX: 0189. H:\HistologyImages\2009_11_16 colon\30347,09\tm\1−158.jpg
IMDBX: 0101. H:\HistologyImages\2009_11_16 colon\31592.09\normal\6x4.jpg
IMDBX: 0189. H:\HistologyImages\2009_11_16 colon\30347,09\tm\1−158.jpg
50
100
150
200
250
300
350
400
450
500
100
200
300
400
500
600
700

(a)

(b)

(c)

(d)

Figure 5.14. QSL labelling results with the ground truth information. Two sample histopathology images with overlaid ground truth atlas data, (a) completely consists of NNCR, (b) has both NNCR and CRCa tissues. QSL labelling results following the ICA (c),(d) . Regions bounded by dashed green lines imply NNCR and solid red lines imply CRCa tissue regions.

CRCa vectors were generated using two different Gaussian distributions separately. The synthetic texture feature vectors were generated for a fixed measure of separation between NNCR and CRCa vector groups. The measure of separation $s_i$ for the $i$'th texture feature vector component was defined as;

$$s_i = \frac{|\mu_{iN} - \mu_{iC}|}{\sigma_{iN}^2 + \sigma_{iC}^2} \tag{5.8}$$

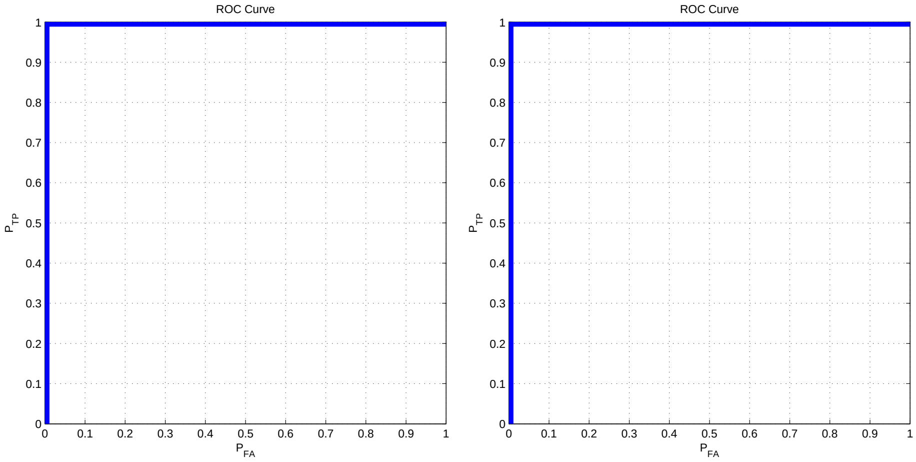
where $\mu_{iN}$ and $\mu_{iC}$ denote the mean values and $\sigma_{iN}$ and $\sigma_{iC}$ the standard deviations of the $i$'th feature vector component for NNCR and CRCa vectors respectively.

In this study, two synthetic texture feature vector datasets were generated by setting $\sigma_{iN} = \sigma_{iC} = 20.0$ and selecting $\mu_{iN}$ and $\mu_{iC}$ so that $s_i$ took values 40.0 and 4.0 respectively. This means that the vector dataset provided with $s_i = 40.0$ was well separated NNCR and CRCa vectors compared to the one provided with $s_i = 4.0$. These synthetic texture feature vectors had the same vector dimensionality and the same number of vectors with the feature vectors of Dataset 5 to provide experimental compatibility (see Table 5.1). Hence, again, the synthetic vector dimensionality was 132 and the number of vectors was 22080 in this case.

The resulting labelling performances for both of the synthetic feature vector datasets were calculated as $(P_{FP}, P_{TP}) = (0.00, 1.00)$ and the ROC Area as 1.0. The ROC curves obtained from the experiments operated on synthetic feature vectors are given in Figure 5.15. These ROC curves, both, represent the ideal labelling case in which the area under the ROC curve equals to 1.0. Therefore, it can be stated that the QSL algorithm could perform very successful classification in case of well separated input texture feature vector distributions.

## 5.9 Segmentation Of The Background Regions

The background regions in histopathology cross section images can be defined as the regions that do not have any stained tissues in the examined specimen. Generally, all background regions are segmented out in the earlier stages of quantitative analysis of histopathology images so that all subsequent operations can focus on the data of interest and any negative effects that may be caused by the background



(a) (b)

Figure 5.15. ROC curves obtained on synthetic feature vector datasets of dimensionality 132 and the number of vectors is 22080. $\sigma_{iN} = \sigma_{iC} = 20.0$ and the measure of separation is (a) 40.0, (b) 4.0.

regions can thus be eliminated.

In the QSL experiments presented so far, the background segmentation was not performed based on the assumption that both the reference and the mixed groups have feature vectors corresponding to the background regions and the QSL algorithm would label these common features as 'normal'. In order to text the validity of this assumption, a simple background segmentation algorithm was implemented and the feature vectors corresponding to the background regions were taken out of the original datasets prior to automated labelling by the QSL algorithm. Finally, the QSL labelling performances obtained from the original texture feature vector datasets and the reduced datasets were compared to each other. The results of these experiments shown below vindicated this assumption; : The removal of the background regions task did not bring any major change to the original QSL labelling performances.

### 5.9.1 Histopathology Image Background Atlas

Histopathology Image Background Atlas is an automatically generated data carrying the local information of being background or foreground for the images in the Histopathological Image Library. Histopathology Image Background Atlas was used to detect background texture feature vectors and hence to remove them from the subsequent analysis. Similar to the Histopathological Image Atlas (see Section 5.2.1), each histopathology image in the Histopathological Image Library was divided into a square grid and each square was automatically labelled as background or foreground. For a grid square to be labelled as background, the sum of local gray level histogram values $h$ corresponding to brighter pixels (equal or greater than gray level a preset $B$) was required to be larger than a predefined threshold $T$ as formulated,

$$\sum_{i=B}^{255} h(i) > T \tag{5.9}$$

where, $i$ stands for the gray level histogram bins. Successful background identification results were obtained when $B = 195$ and $T = 0.95$ as determined experimentally. The background grid square size was 256 microns. The total number and the percentages of background grid squares calculated for the images in the Histopathological Image Library are given in Table 5.12. In addition, the total number of background grid squares and the percentages to all available squares for both the reference and mixed image groups are also listed. It can be observed that the background region percentage is higher in the reference group (NNCR) compared to the mixed image group (NNCR + CrCa).

Table 5.12. The total number and percentages of background squares (of square grid dimension 256 microns) in the Histopathology Image Background Atlas for the reference (NNCR), the mixed groups (NNCR + CrCa) and total.

| Images | Total | Background | Percentage (%) |
|---|---|---|---|
| Reference | 22860 | 1507 | 6.59 |
| Mixed | 18540 | 299 | 1.61 |
| All | 41400 | 1806 | 4.36 |

Several examples of background segmented images from the Histopathological Image Library are given in Figure 5.16 with both the background and the foreground regions onto each image.

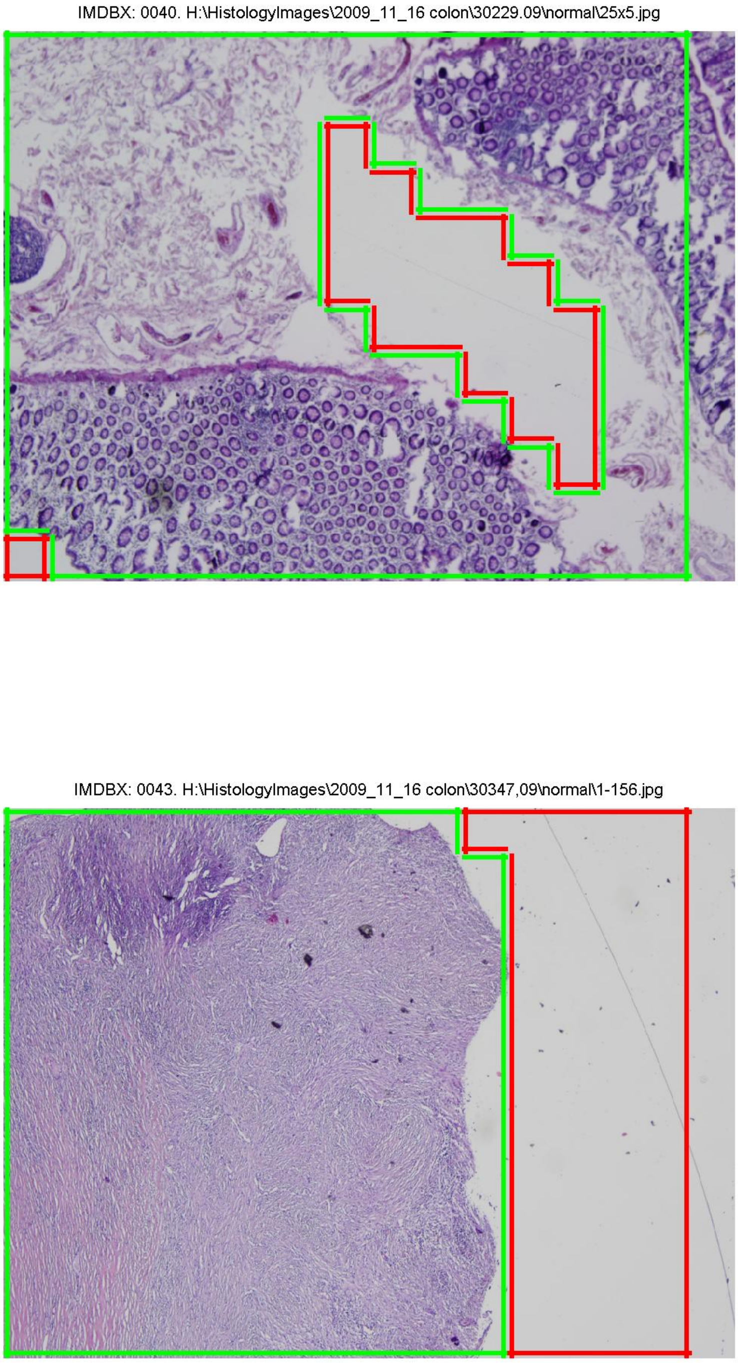

IMDBX: 0040. H:\HistologyImages\2009_11_16 colon\30229.09\normal\25x5.jpg
IMDBX: 0043. H:\HistologyImages\2009_11_16 colon\30347,09\normal\1-156.jpg

Figure 5.16. Background segmented images by the proposed segmentation method. Regions bounded by green lines imply foreground and red lines imply background regions.

Several texture labelling experiments were performed in order to assess how the removal of the background regions affected the labelling performance of the QSL method. The original texture feature datasets and the segmented texture feature vector datasets were used as QSL inputs and the resulting ROC curves were compared. In the baseline experiment which operating on texture feature vector Dataset 5, the number of texture feature vectors was 22080. The secondary vector dataset had 20836 texture feature vectors due to the removal of 1244 vectors by the background segmentation. The resulting ROC curves for the baseline and the background segmented experiments were very close to each other (see Figure 5.17). This suggests that the proposed methodology does not require any preliminary background segmentation tasks.

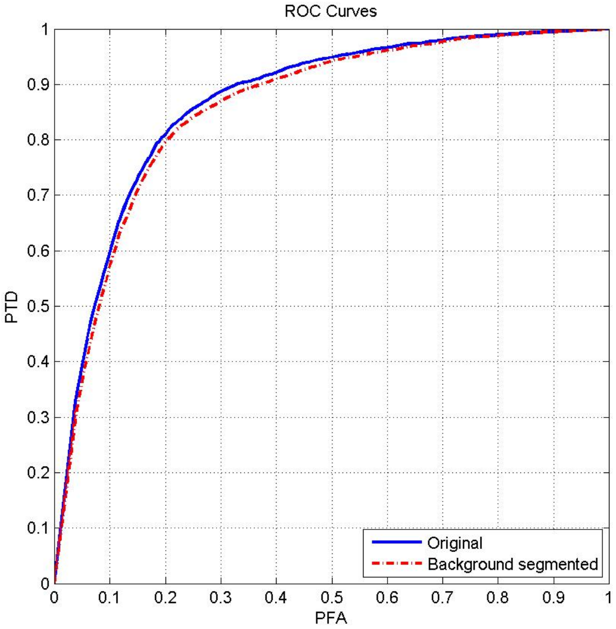


Figure 5.17. ROC curves for the original and background segmented texture feature vectors (Dataset 5).

## 5.10 Target Texture Classification By The Quasi-supervised Learning Algorithm

This section describes the predictive classification of the target texture vectors following the QSL labelling phase. In target texture classification experiments, a small portion of the Histopathological Image Library images were reserved for testing purposes; 27 of them were selected against 230 training images.

The QSL predictive classifier used $n_{opt}$ values calculated in the prior QSL labelling experiments of the corresponding feature vector datasets (see Sections 4.2.3 and 5.4.1). The QSL classifier computed the posterior probabilities $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$ for all feature vectors in a target vector dataset for the optimal reference set size $n_{opt}$ with respect to the texture feature vectors from the training images. Using these $p_0(\mathbf{x})$ values, the ROC curves were plotted and the optimum threshold $T_{opt}$ value was determined. Subsequently, using this $T_{opt}$ value, corresponding resulting texture labels were determined and the texture recognition performances were calculated by comparing the results with the ground truth labels of the Histopathological Image Atlas.

The labels of the regions in the target images were determined in accordance with the labels of their texture feature vectors predicted by the QSL algorithm.

### 5.10.1 Performance Evaluation For The Target Texture Classification

In order to assess the separation of the NNCR and CRCa tissue regions in a target texture vector dataset, the receiver operating characteristics (ROC) curves were constructed. Similar to the QSL labelling case, the false positive rates $P_{FP}$ were computed on the corresponding training reference vector dataset. However, $P_{TP}$ values were not calculated on the unlabelled mixed vector dataset this time; instead, the probability of detection ($P_D$), in other words, probability of positives were calculated on the target vector dataset. Thus, this version of the ROC curves was changed to a graphical plot of the positive detection rate for the target vector group versus the false positive rate for the reference vector group. To generate an

ROC curve, we have computed $P_{FP}$ on the reference vector dataset and $P_D$ on the target vector dataset and plotted both for varying threshold values of $T$ ranging from 0.0 to 1.0. Similarly, the $T$ value that corresponds to the knee point of the continuous ROC curve where the tangent equals to 1.0 was selected as the optimum threshold value $T_{opt}$.

The ROC curves obtained by the QSL target classifier operated on Datasets 1-6 are given in Figure 5.18. According to the performance comparison rules described in Section 5.2 it was observed that the worst texture recognition performance was acquired with texture vector Dataset 1 corresponding to the gray level information (see Table 5.1 ).

The QSL target recognition performances for texture feature datasets, Datasets 1-6 are presented in Table 5.13 in which the optimum posterior probability threshold values $T_{opt}$ and the areas under the ROC curves are listed with the resultant true positive and false positives calculated using the ground truth Histopathological Image Atlas labels. It was observed that according to the "ROC Area" criterion, the worst recognition rate was obtained for Dataset 1 which is based on gray level image information.

In general, the false positive rates are quite high, though the true positive rates reach up to 92%. This phenomenon is referred to the "malign tendency" and was also observed in the analytical labelling experiments as explained in Section 5.4.2.

Two QSL target classification results as the resultant vector class labels overlaid onto histopathology images are presented in Figure 5.19 and Figure 5.20. These figures are the histopathology images with the ground truth atlas data and the resultant QSL classification labels overlaid onto original images. Figure 5.19 has two histopathology images consisting of only NNCR tissues. The QSL classifier results with predicted CRCa labels as shown in Figure 5.19(b) and Figure 5.19(d), clearly, are false positive results.

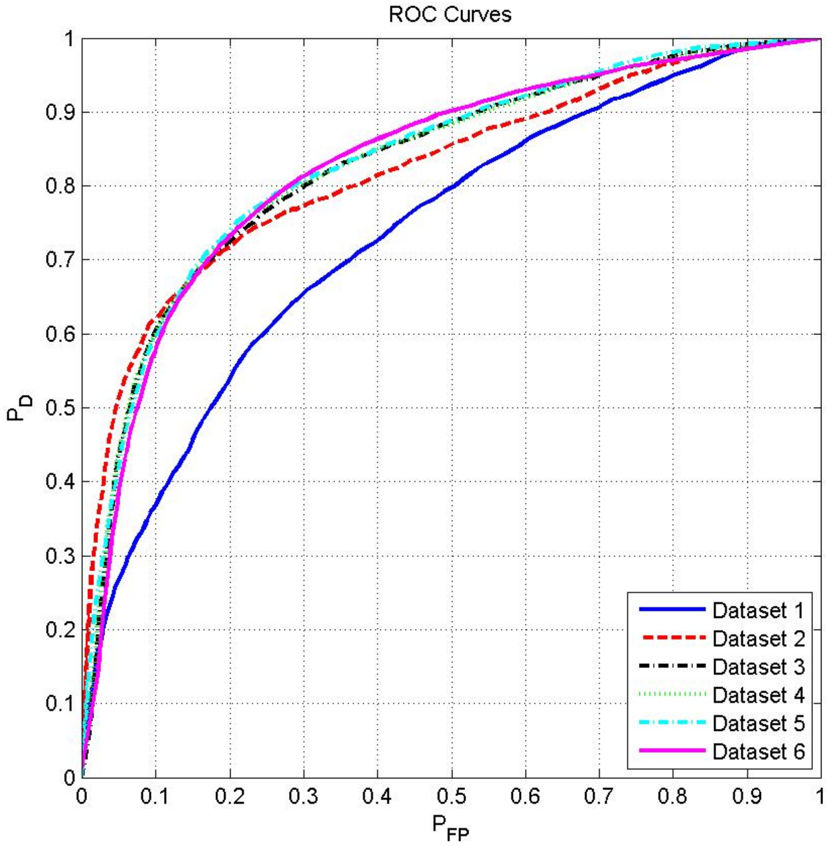
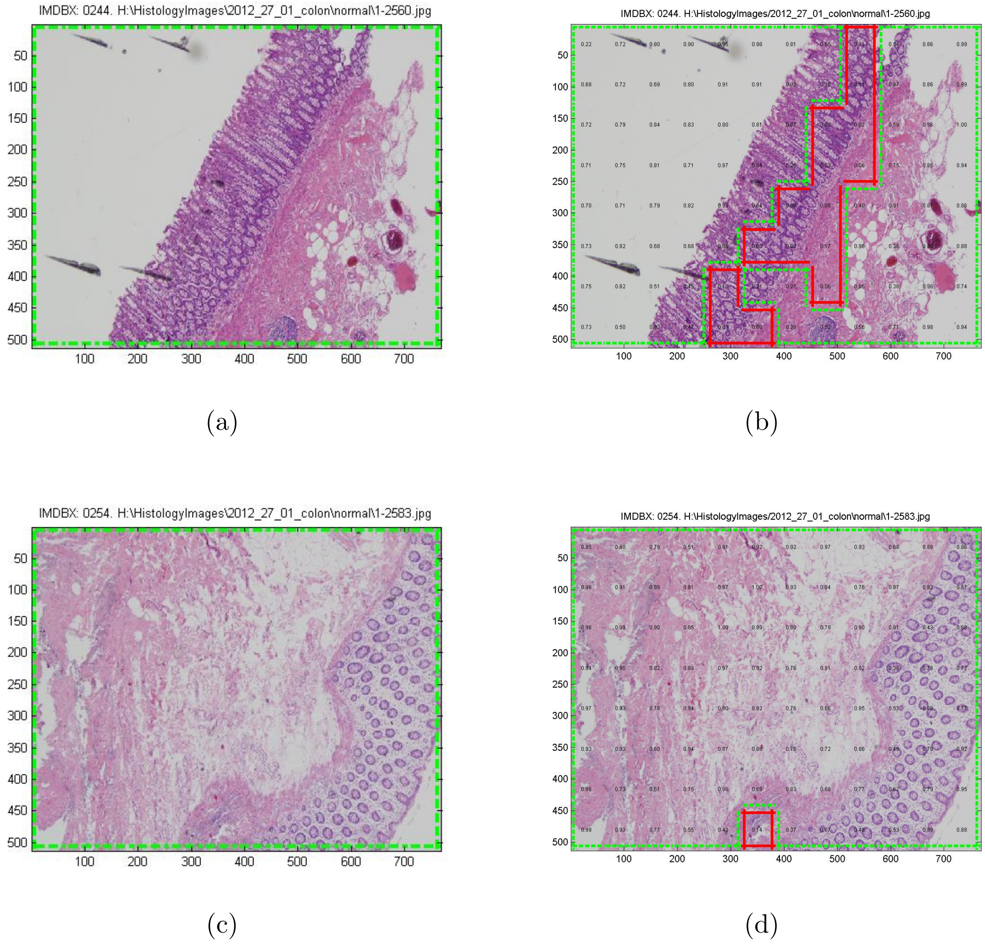


Figure 5.18. ROC curves obtained for the predictive classification of the target texture feature vectors (Dataset 1-6). Positive detection rate $P_D$ for the target vector groups are plotted versus the false positive rate $P_{FP}$ for the training reference vector group.


IMDBX: 0244. H:\HistologyImages\2012_27_01_colon\normal\1-2560.jpg
IMDBX: 0244. H:\HistologyImages/2012_27_01_colon\normal\1-2560.jpg
IMDBX: 0254. H:\HistologyImages\2012_27_01_colon\normal\1-2583.jpg
IMDBX: 0254. H:\HistologyImages/2012_27_01_colon\normal\1-2583.jpg
50
100
150
200
250
300
350
400
450
500
100
200
300
400
500
600
700

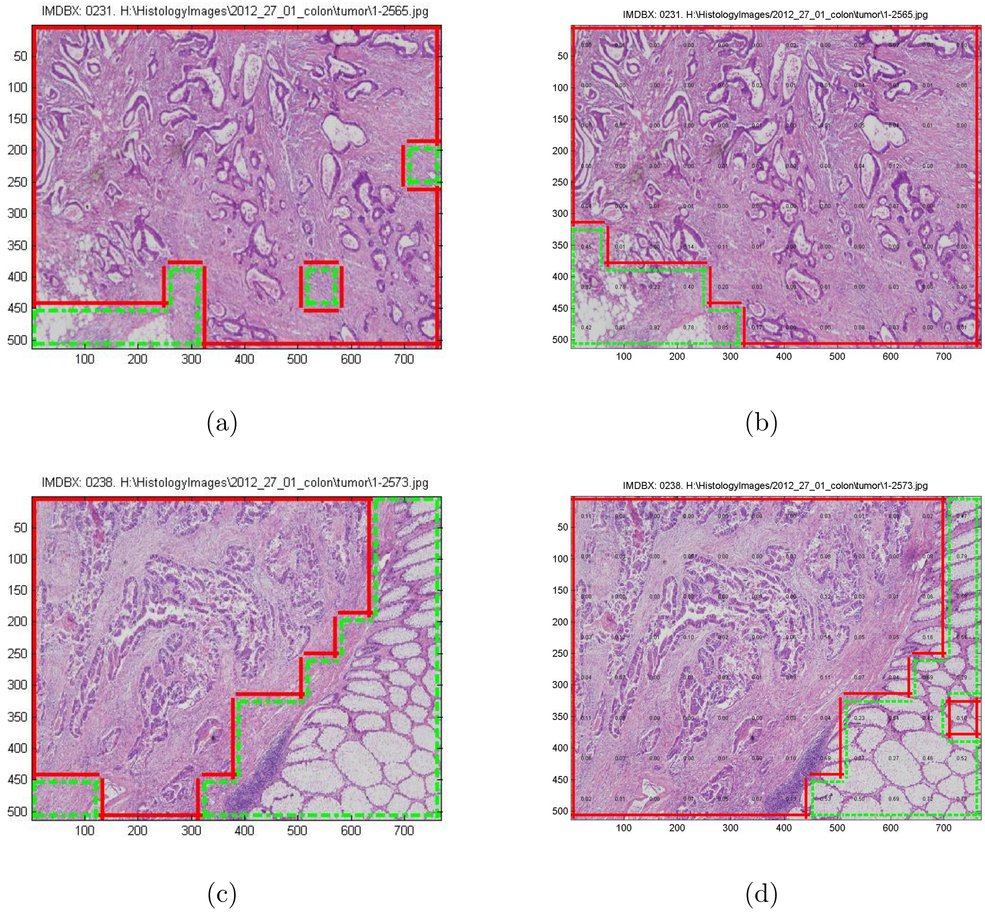
(a)

(b)

(c)

(d)

Figure 5.19. QSL target classification results along with the ground truth information. Two histopathology images completely consisting of NNCR tissues, with overlaid ground truth atlas data (a), (c). The same histopathology images with overlaid QSL target classification results (b), (d). Regions bounded by dashed green lines imply NNCR and solid red lines imply CRCa tissue regions.


IMDBX: 0231. H:\HistologyImages\2012_27_01_colon\tumor\1-2565.jpg
IMDBX: 0231. H:\HistologyImages/2012_27_01_colon\tumor\1-2565.jpg
IMDBX: 0238. H:\HistologyImages\2012_27_01_colon\tumor\1-2573.jpg
IMDBX: 0238. H:\HistologyImages/2012_27_01_colon\tumor\1-2573.jpg
50
100
150
200
250
300
350
400
450
500
100
200
300
400
500
600
700

(a)

(b)

(c)

(d)

Figure 5.20. QSL target classification results along with the ground truth information. Histopathology images having both NNCR and CRCa tissues, with overlaid ground truth atlas data (a), (c). QSL target classification results overlaid (b), (d). Regions bounded by dashed green lines imply NNCR and solid red lines imply CRCa tissue regions.

## 5.11 Comparative Target Texture Classification By The Support Vector Machine Classifier

In order to obtain an independent evaluation of the QSL target texture recognition performance, a Support Vector Machine (SVM) classifier was used. The SVM classifier was trained on the ground-truth label data that was withheld from the quasi-supervised labelling strategy similar to the comparison of training performances as explained in Section 5.6. In order to perform the SVM target classification experiments, the texture vector datasets were divided into training and target sets. The target texture vectors were extracted from the fixed 27 target HIL images. The rest of the HIL images were left as the training images, hence, as training texture vector source. In order to set up the SVM training setup, we constructed both the NNCR and CRCa vector groups using ground-truth atlas information. The group vectors took labels $+1$ and $-1$, regarding to the condition if they belong to NNCR group or not.

The training feature vector sets were used to obtain the SVM classifier model, and this classifier model was applied to the target feature vector set. The SVM classifier output obtained for the vectors in the target datasets was then compared with the ground truth labels in the Histopathological Image Atlas to calculate the resultant classifier performances.

The SVM classifier target recognition performances for texture feature Datasets 1-6 are presented along with the QSL performance values in Table 5.13. In the table, the resultant true positive and false positive rates were also listed.

The worst $P_{FP}$ value was obtained for Dataset 1 but with the highest $P_{TP}$ value of all. For all of the texture vector datasets, the SVM classifier resulted with lower $P_{FP}$ values compared to the QSL. However, the QSL had larger $P_{TP}$ values except for the value of Dataset 1 which was very close to that of the SVM. It should be kept in mind, however, that the QSL algorithm produced predictions without the benefit of a ground-truth dataset from the atlas in contrast to the SVM classifier.

Table 5.13. Target texture recognition performances of the QSL and the SVM classifier.

| | QSL | | | SVM |
|---|---|---|---|---|
| Texture Features | $T_{opt}$ | ROC Area | $(P_{FP}, P_{TP})$ | $(P_{FP}, P_{TP})$ |
| Dataset 1 | 0.53 | 0.74 | (0.50, 0.86) | (0.21, 0.90) |
| Dataset 2 | 0.41 | 0.82 | (0.54, 0.89) | (0.09, 0.75) |
| Dataset 3 | 0.48 | 0.83 | (0.60, 0.91) | (0.10, 0.79) |
| Dataset 4 | 0.51 | 0.83 | (0.62, 0.92) | (0.13, 0.79) |
| Dataset 5 | 0.49 | 0.83 | (0.65, 0.89) | (0.16, 0.73) |
| Dataset 6 | 0.55 | 0.83 | (0.65, 0.92) | (0.11, 0.76) |

## 5.12 Target Texture Classification And Dimensionality Reduction By The Independent Component Analysis

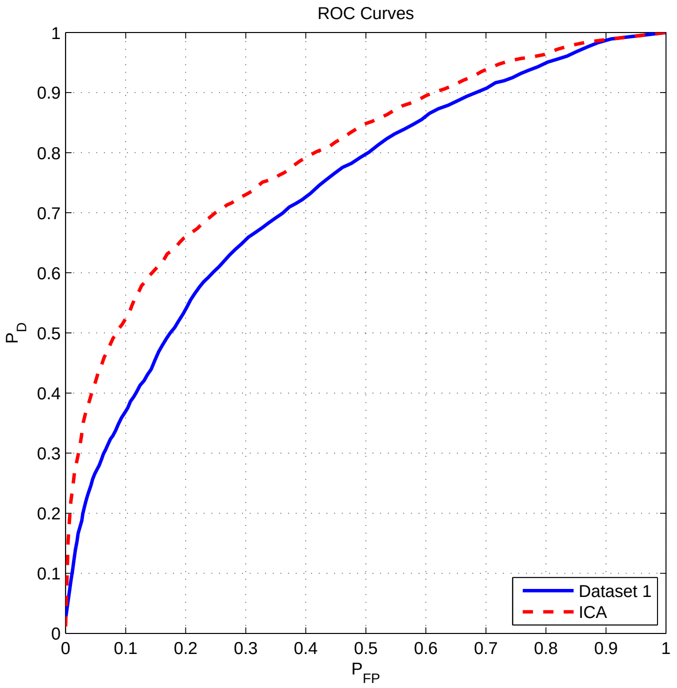
In these experiments, the original texture feature vector datasets were converted to reduced dimensionality feature vector datasets using the FastICA method (see Section 4.4.8). The reduced vector datasets were then fed to the QSL algorithm and the results generated were compared to the results of the baseline experiments.

The classification performance measures for the baseline and the ICA applied experiments on Dataset 1-5 are presented in Table 5.14. The results indicated that for Dataset 1 and 5, the ICA method improved the classification performances; increased the areas under the ROC curves and the optimum recognition points on the ROC curves moved closer towards the ideal point.

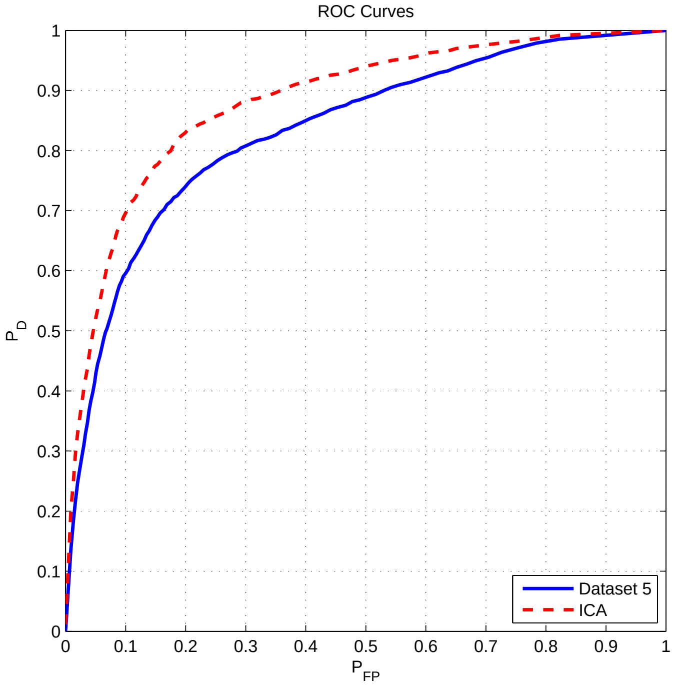
A couple of ROC curves obtained from the baseline and the ICA-applied target classification experiments are given in Figure 5.21. It can be observed that the ROC curves obtained from the ICA applied vector datasets resulted with higher classification performances compared to the baseline ROC curves.

Table 5.14. Target classification performances of the QSL method for original and the ICA applied texture feature vector datasets.

| | Original | | ICA | |
|---|---|---|---|---|
| Texture Features | $(P_{FP}, P_{TP})$ | ROC Area | $(P_{FP}, P_{TP})$ | ROC Area |
| Dataset 1 | (0.50, 0.86) | 0.74 | (0.48, 0.95) | 0.80 |
| Dataset 2 | (0.54, 0.89) | 0.82 | (0.35, 0.88) | 0.74 |
| Dataset 3 | (0.60, 0.91) | 0.83 | (0.48, 0.95) | 0.75 |
| Dataset 4 | (0.62, 0.92) | 0.83 | (0.28, 0.78) | 0.64 |
| Dataset 5 | (0.65, 0.89) | 0.83 | (0.75, 0.96) | 0.88 |



(a)



(b)

Figure 5.21. ROC curve pairs for the baseline and the ICA applied vector datasets. The baseline and the reduced vector dimensions for (a) Dataset1 are 296 and 279, (b) Dataset5 are 132 and 119.

# CHAPTER 6

# CONCLUSION

## 6.1 Summary

This thesis presents an evaluation of a quasi-supervised learning methodology in quantitative histopathology image analysis and biomedical feature data classification.

Experimental results in Section 5.6 has showed that, the QSL method achieved satisfactory accuracy levels in texture labelling. Since there is no histopathology image texture benchmark database providing a comparison of accuracies achieved by other classification schemes, it was unfortunately not possible to check our results with the other methods from the literature. Yet, in experiments against an SVM classifier using a ground-truth training dataset lacking in a quasi-supervised setting, the QSL method proved itself with accuracy levels close to the upper bounds provided by the idealized SVM classification. The probability of false positive values by the SVM classifier were relatively lower compared to those by the QSL algorithm. On the other hand, the probability of true positive values were very close for both methods.

In Section 5.3, we have described the computation of several different texture feature vector datasets. The texture features were calculated in a directionally invariant manner. To this end, we have derived the first order texture features using the local color histograms and the second order texture features from co-occurrence matrices. The different feature vector datasets corresponding to different feature extraction configurations allowed us to perform many QSL experiments and to compare the resulting labelling performances.

Among the various datasets corresponding to different feature extraction parameters, the labelling performance of the QSL method was found the poorest when operated on the dataset constructed using the gray level image information. The color texture feature information derived from the *Lab* color space offered the best characterization of the NNCR and CRCa features.

It was also observed that using multiple-scale feature vectors did not have any significant influence on the QSL texture labelling performance. The varying values of the hierarchical computation scale parameter $H$ did not have any significant impact on the QSL texture labelling performance either, only a minor performance improvement of true positive values was observed for $H = 2$ at the optimum performance points.

In the labelling performance comparisons, the QSL labelling performances for texture feature vector datasets calculated using a range of inter-pixel distances $d$ were better than the one calculated for only $d = 1$. This means that combined texture features from several inter-pixel distance values improved the separation between healthy and cancer texture features vectors.

In this study, we have experimented with several dimensionality reduction procedures in Section 5.7, to determine if reducing the texture vector dimensionalities would lead better classification performances. Among these procedures, only the ICA method improved the labelling performance. One of the possible reasons for the improvement is that ICA could extract the valuable vector component information despite their low variances and also suppresses the redundant data. In the other dimensionality reduction procedures, Individual Feature Selection, Principal Component Analysis, and Isomap, the labelling accuracy of the baseline experiments on the original datasets were higher than the reduced dimensionality vector datasets. For all of these methods, the labelling accuracy started with lower values for smaller target dimensions, and as target dimensionality increased, the accuracy levels moved closer to that of the baseline experiment but never exceeded it. This behaviour implies that the QSL method was robust to the Curse Of Dimensionality phenomenon and this was a very profound contribution of this study.

Next, we have evaluated the first order and the second order texture features separately in Section 5.7.1. In experiment results, the ROC curves obtained from the second order feature characteristics were above the ones obtained from the first order feature characteristics. On the other hand, the ROC curves obtained from the second order feature characteristics alone were bounded by the original ROC curves.

This means that better labelling performances were reached by concatenating both the first order and the second order texture feature vector subsets in comparison to the performances obtained by using them individually.

We have then performed a background segmentation algorithm to take feature vectors corresponding to the background regions out of the analysis as explained in Section 5.9. The background regions correspond to transparent parts of the histopathology slides which contain no valuable textural information. In experiments, the background segmentation task did not bring any major change to original QSL labelling performances. This result reinforces that the proposed methodology does not need any pre-requisite background segmentation task because, in essence, both the reference and the test groups are expected to have almost equal amounts of background regions.

Finally, we have evaluated the predictive behaviour of the QSL via several target classification experiments in Section 5.10. Similar to the texture labelling case, we ave performed performance comparison with the SVM classifier using a ground-truth training dataset. The probability of false positive values by the SVM classifier were relatively lower compared to those by the QSL algorithm. On the other hand, the probability of true positive values by the QSL algorithm were higher than that of SVM classifier except the experiment operated on gray level texture information. In addition, we have experimented with the ICA dimensionality reduction procedure to determine if reducing the texture vector dimensionalities would lead better target classification performances compared to the baseline experiments as observed in most of the texture labelling experiments. The result is that, in some of the experiments, the ICA method improved the classification performances as presented in Section 5.12.

## 6.2 Future Study

Despite the high performance levels obtained in the form of true positives, we have faced a phenomenon referred to as “malign tendency ” which came up with high false positive rates, implying a general bias towards labelling NNCR texture vectors as CRCa, observed especially in the mixed labelled test group. We believe that this phenomenon is due to severely complicated local structures in colonic tissues and the resulting weak separation of normal and cancer feature vectors in the multidimensional vector space. Moreover, we have checked this assumption by performing several QSL experiments on synthetic vector data. We have generated statistically well differentiated random feature vectors for both NNCR and CRCa tissues in high dimensional space and obtained excellent QSL labelling performances. This evidence showed us that in order to increase the QSL accuracy, different texture feature characteristics must be incorporated into the automated labelling framework. A feature extraction scheme using co-occurrence matrices calculated by the pixel pairs of a specified orientation could be tested. There are also several alternatives to the features used in this study as described in Section 3.1.2 and the comparison of their labelling performances with the QSL accuracies remains to be evaluated for the optimum configuration. Some examples to alternative texture features can be listed as;

- Fractals,
- Gabor based features,
- Morphological features,
- Zernike moments,
- Fourier spectral characteristics.

As stated before, the histopathological slides used in our experiments were well balanced in terms of the stains absorbed by tissues and the images in the Histopathological Image Library (HIL) were acquired using fixed capture and illumination parameters. Thus, our assumption was that these images were assumed to share similar visual standards and lack artefacts that can occur due to variations in the sample preparation or image acquisition procedures. However, whether these images possessed a variation imperceptible to the naked eye in the H&E staining process that might be affecting the computerised analysis remains an open question. Another research can thus be carried out by performing a color normalisation method in the early stage of the proposed framework using one of the methods described in Section 3.1.1 prior to statistical analysis using the QSL method.

In this research, many dimensionality reduction procedures were experimented and the ICA method was found to be the only one improving the labelling accuracy. However, several other feature selection or extraction algorithms leading to dimensionality reduction could not be evaluated due to long computation times required by these methods for the texture feature vector datasets used in this study. In the future, these algorithms, such as Exhaustive Selection and Branch and Bound method described in Section 4.4.3 can be applied to the existing feature vector datasets provided that they can be carried out within limits of computational feasibility.

# REFERENCES

Aizerman, A., Braverman, E. M., Rozoner, L. I., 1964. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25, 821–837.

Albert, R., Schindewolf, T., Baumann, I., Harms, H., 1992. Three-dimensional image processing for morphometric analysis of epithelium sections. Cytometry 13, 759–765.

Bellman, R. E., 1961. Adaptive control processes: a guided tour. Princeton University Press.

Bishop, C. M., Svensn, M., Williams, C. K. I., 1998. The generative topographic mapping. Neural Computation, 215–234.

Boser, B. E., Guyon, I. M., Vapnik, V., 1992. A training algorithm for optimal margin classifiers. In: . Fifth Annual Workshop on Computational Learning Theory, pp. 144–152.

Boucheron, L. E., 2008. Object and spatial level quantitative analysis of multispectral histopathology images for detection and characterization of cancer. Ph.D. thesis, Univ. of California, Santa Barbara, CA.

Burges, J. C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery.

Burt, P., Adelson, E., 1983. The laplacian pyramid as a compact image code. IEEE Transactions on Communications 31, 532–540.

Can, A., Bello, M., Cline, H. E., Tao, X., Ginty, F., Sood, A., Gerdes, M., Montalto, M., 2008. Multi-modal imaging of histological tissue sections. In: Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on. pp. 288–291.

Chapelle, O., Schölkopf, B., Zien, A., 2006. Introduction to Semi-Supervised Learning. The MIT Press.

CIE, 1986. CIE colorimetry: Official recommendations of the international commission on illumination. Tech. rep., CIE.

Comon, P., 1994. Independent Component Analysis: a new concept ? Signal Processing.

Cortes, C., Vapnik, V. N., 1995. Support vector networks. Machine Learning.

Cover, T. M., Campenhout, J. M. V., 1977. On the possible orderings in the measurement selection problem. IEEE Transactions On Systems, Man, and Cybernetics 7 (9), 657–661.

Cover, T. M., Hart, P. E., 1967. Nearest neighbor pattern classification. Information Theory.

Cover, T. M., Thomas, J. A., 1991. Elements of Information Theory. Wiley-Interscience.

Demir, C., Yener, B., 2006. Automated cancer diagnosis based on histopathological images: A systematic survey. Tech. rep., Rensselaer Polytechnic Institute, Troy, NY.

Diamond, D. A., Berry, S. J., Umbricht, C., Jewett, H. J., Coffey, D. S., 1982. Computerised image analysis of nuclear shape as a prognostic factor for prostatic cancer. Prostate 3, 321–332.

Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M., Tomaszeweski, J., April 2007. Automated grading of prostate cancer using architectural and textural image features. In: Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on. No. 9506433. Dept. of Biomed. Eng., State Univ. of New Jersey, New Brunswick, NJ, pp. 1284 – 1287.

Doyle, S., Rodriguez, C., Madabhushi, A., Tomaszeweski, J., Feldman, M., 2006. Detecting prostatic adenocarcinoma from digitized histology using a multi-scale hierarchical classification approach. In: Engineering in Medicine and Biology Society. EMBS '06. 28th Annual International Conference of the IEEE. pp. 4759 – 4762.

Duda, R. O., Hart, P. E., Stork, D. G., 2000. Pattern Classification. Wiley-Interscience.

Esgiar, A., Naguib, R., Bennett, M., Murray, A., AUG 1998a. Automated feature extraction and identification of colon carcinoma. Analytical And Quantitative Cytology And Histology 20 (4), 297–301, 11th International Congress on Diagnostic Quantitative Pathology, Siena, Italy, Oct 02-04, 1997.

Esgiar, A. N., Naguib, R. N. G., Sharif, B. S., Bennett, M. K., Murray, A., March 2002. Fractal analysis in the detection of colonic cancer images. IEEE Transactions on Information Technology in Biomedicine 6.

Esgiar, N., Naguib, A., Sharif, R. N. G., Bennett, B. S., M.K. Murray, A., 1998b. 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.

Ficsor, L., Varga, V., Tagscherer, A., Tulassay, Z., Molnar, B., Mar 2008. Automated classification of inflammation in colon histological sections based on digital microscopy and advanced image analysis. Cytometry 73 (3), 230–7.

Filippas, J., Amin, S., Naguib, R., Bennett, M., 2003. A parallel implementation of a genetic algorithm for colonic tissue image classification. In: Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topic Conference on. pp. 330 – 333.

Fox, H., 2000. Is H&E morphology coming to an end ? Journal Of Clinical Pathology 53, 38–40.

Friedrich, T., 2002. Nonlinear dimensionality reduction with locally linear embedding and Isomap. Master's thesis, Department of Computer Science The University of Sheffield.

Fukunaga, K., 1982. Intrinsic dimensionality extraction. Vol. 2 of Classification, Pattern Recognition and Reduction of Dimensionality, of Handbook of Statistics. North Holland.

Fukunaga, K., Hostetler, L. D., 1975. k-nearest-neighbor Bayes-risk estimation. IEEE Transactions on Information Theory 21 (3), 285–293.

Gonzalez, R., Woods, R., 1992. Digital Image Processing. Addison-Wesley.

Gurcan, M., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., Yener, B., 2009. Histopathological image analysis: A review. IEEE Reviews in Biomedical Engineering 2, 147–171.

Güven, M., 2010. Detection of man-made structures in aerial imagery using quasi-supervised learning and texture features. Master's thesis, İzmir Institute of Technology.

Hallouche, F., Adams, A. E., Hinton, O. R., Relf, G., Lakshmi, M. S., Sherbet, G. V., 1992. Image processing for cell cycle analysis and discrimination in metastatic variant cell lines of B16 murine melanoma. Pathobiology 60, 76–81.

Hamilton, Allen, D. C., Watt, P. C. H., 1987. Classification of normal colorectal mucosa and adenocarcinoma by morphometry. Histopathology 11, 901–911.

Hamilton, P. W., Bartels, P., Thompson, D., Anderson, N. H., Montironi, R., Sloan, J., May 1997. Automated location of dysplastic fields in colorectal histology using image texture analysis. Journal Of Pathology 182 (1), 68–75.

Haralick, R. M., Shanmugam, K., Dinstein, I., 1973. Textural features for image classification. IEEE Transactions On Systems, Man, And Cybernetics.

Haykin, S., 2008. Neural Networks and Learning Machines, 3rd Edition. Prentice-Hall.

Haykin, S., Chen, Z., 2005. The cocktail party problem. Neural Computation.

Henze, N., Penrose, M., 1999. On the multivariate runs test. In: Annals of Statistics. pp. 290–298.

Huber, P., 1985. Projection pursuit. In: The Annals of Statistics. Vol. 13. pp. 435–475.

Hyvärinen, A., 1999. Fast and robust fixed-point algorithms for Independent Component Analysis. IEEE Transactions on Neural Networks 10 (3), 626–634.

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis.

Hyvärinen, A., Oja, E., 1997. A fast fixed-point algorithm for Independent Component Analysis. Neural Computation 9 (7), 1483–1492.

Hyvärinen, A., Oja, E., 2000. Independent Component Analysis: Algorithms and applications. Neural Networks, 411–430.

Jafari-Khouzani, K., Soltanian-Zadeh, H., 2003. Multiwavelet grading of pathological images of prostate. IEEE Transactions On Biomedical Engineering.

Jain, A., Zongker, D., 1997. Feature selection: Evaluation, application, and small sample performance. IEEE Transaction On Pattern Analysis And Machine Intelligence 19, 153–158.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. In: European Conference on Machine Learning (ECML). Springer, Berlin, pp. 137–142.

Joachims, T., 1999a. Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning. Vol. 11. MIT Press.

Joachims, T., 1999b. Transductive inference for text classification using support vector machines. In: International Conference on Machine Learning (ICML). Bled, Slowenien, pp. 200–209.

Joachims, T., 2002. Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms. Kluwer/Springer.

Jolliffe, I., 2002. Principal Component Analysis.

Karaçalı, B., 2010. Quasi-supervised learning for biomedical data analysis. Pattern Recognition.

Karaçalı, B., Krim, H., 2003. Fast minimization of structural risk by nearest neighbor method. IEEE Transactions on Neural Networks 14 (1), 127–137.

Karaçalı, B., Ramanath, R., Snyder, W., 2004. Structural risk minimization-based nearest neighbor classifier. Pattern Recognition Letters 25 (1), 63–71.

Keenan, S., Diamond, J., McCluggage, W. G., Bharucha, H., Thompson, D., Bartels, P., Hamilton, P., 2000. An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN). Journal of Pathology 192 (3), 351–362.

Kinzler, K. W., Vogelstein, B., October 1996. Lessons from hereditary colorectal cancer. Cell 87, 159–170.

Kittler, J., 1978. Feature set search algorithms. Pattern Recognition and Signal Processing, 41–60.

Kohonen, T., September 1990. The self organizing map. Proceedings of the IEEE Transactions on Computers, 1464–1480.

Köktürk, B. E., 2011. Separation of stimulus-specific patterns in electroencephalography data using quasi-supervised learning. Master's thesis, İzmir Institute of Technology.

Kong, J., Sertel, O., Shimada, H., Boyer, K., Saltz, J., Gurcan, M., 2009. Computer-aided evaluation of neuroblatoma on whole-slide histology images: Classifying grade of neuroblastic differentiation. Pattern Recognition 42, 1080–1092.

Kruskal, J. B., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29, 1–27.

Kumar, V., Grama, A., Gupta, A., Karypis, G., November 1993. Introduction to Parallel Computing: Design and Analysis of Algorithms. Benjamin-Cummings Publishing Company.

Lanza, G., Messerini, L., Gafc, R., Risio, M., 2011. Colorectal tumors: The histology report. Digestive and Liver Disease, 344–355.

Lee, J. A., Lendasse, A., Verleysen, M., April 2002. Curvilinear distance analysis versus Isomap. In: Proceedings of ESANN2002, 10th European Symposium on Artificial Neural Networks. pp. 185–192.

Luenberger, D., 1969. Optimization by Vector Space Methods. Wiley.

Macenko, M., Niethammer, M., Marron, J., Borland, D., Woosley, J., Guan, X., Schmitt, C., Thomas, N., 2009. A method for normalizing histology slides for quantitiative analysis. In: Proceedings of the International Symposium on Biomedical Imaging (ISBI). p. 1107–1110.

Magee, D., Treanor, D., Crellin, D., Shires, M., Mohee, K., Quirke, P., 2009. Colour normalisation in digital histopathology images.

Mardia, K. V., Kent, J. T., Bibby, J., 1995. Multivariate Analysis. Probability and Mathematical Statistics. Academic Press.

Masood, K., Rajpoot, N., Qureshi, H., 2006. Co-occurrence and morphological analysis for colon tissue biopsy classification. In: Proceedings 4th International Workshop on Frontiers of Information Technology (FIT06), Islamabad.

McLachlan, G. J., 2004. Discriminant Analysis and Statistical Pattern Recognition. Wiley Seriesin Probability and Statistics. Wiley-Interscience.

Medical, N., May 2012. Colorectal cancer diagnosis.
URL `http://www.news-medical.net/health/Colorectal-Cancer-Diagnosis.aspx`

Mendez, A. J., Tahoces, P. G., Lado, M. J., Souto, M., Vidal, J. J., June 1998. Computer-aided diagnosis: Automatic detection of malignant masses in digitized mammograms. Medical Physics 25, 957–964.

Naik, S., Doyle, S., Madabhushi, A., Tomaszeweski, J., Feldman, M., 2007. Automated gland segmentation and gleason grading of prostate histology by integrating low, high level and domain specific information. Workshop on Microscopic Image Analysis With Applications in Biology.

Narendra, P. M., Fukunaga, K., Sept 1977. A branch and bound algorithm for feature subset selection. IEEE Transactions on Computers 26 (9), 917–922.

Nasir, K. R., Rajpoot, K. M., Rajpoot, N. M., Turner, M. J., 2004. Hyperspectral colon tissue cell classification.

Neemuchwala, H., Hero, A. O., 2005. Entropic graphs for registration. In: R.S.Blum, Z.Liu (Eds.), Multi-Sensor Image Fusion and its Applications. Marcel Dekker, pp. 185–235.

Notes On Gastrointestinal Histology, University of Ottawa, 2012. The large intestine, gastrointestinal histology.
URL http://www.courseweb.uottawa.ca/medicine-histology/english/gastrointestinal/Gastro_Large_Intest.htm

Nwoye, E., Khor, L. C., Woo, W. L., Dlay, S. S., 2006. Spectral and statistical features in fuzzy neural expert machine for colorectal adenomas and adenocarcinoma classification. In: 5th International Symposium on Communication Systems, Networks and Digital Signal Processing. pp. 792–796.

Önder, D., Karaçalı, B., 2009. Automated clustering of histology slide texture using co-occurrence based grayscale image features and manifold learning. In: Biomedical Engineering Meeting, 2009. BIYOMUT 2009.

Önder, D., Sarıoğlu, S., Karaçalı, B., 2010. Automated classification of cancerous textures in histology images using quasi-supervised learning algorithm. In: Biomedical Engineering Meeting, 2010. BIYOMUT 2010.

Papoulis, A., 1991. Probability, Random Variables and Stochastic Processes. Mcgraw-Hill College.

Pham, D. T., Garrat, P., Jutten, C., 1992. Separation of a mixture of independent sources through a maximum likelihood approach. In: Proc. EUSIPCO. pp. 771–774.

Pitts, D. E., Premkumar, Saganti, B., Houston, A. G., Babaian, R. J., Troncoso, P., 1993. Texture analysis of digitized prostate pathologic cross section. In: Proc. SPIE: Med. Imaging: Image Processing. Vol. 1898. SPIE, pp. 465–470.

Pratt, R., 1991. Digital Image Processing. New York: John Wiley.

Qureshi, H., Sertel, O., Rajpoot, N., Wilson, R., Gurcan, M. N., 2008. Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification. In: MICCAI (2)'08. pp. 196–204.

Rabinovich, A., Laris, C. A., Agarwal, S., Price, J. H., Belongie, S., 2004. Unsupervised color decomposition of histologically stained tissue samples. In: Advances in Neural Information Processing Systems. MIT Press.

Rajpoot, K., Rajpoot, N., 2004. SVM optimization for hyperspectral colon tissue cell classification. In: MICCAI (2). p. 829.

Reinhard, E., Ashikhmin, M., Gooch, B., Shirley, P., September 2001. Color transfer between images. IEEE Computer Graphics and Applications 21 (5), 34–41.

Roweis, S. T., Saul, L. K., December 2000. Nonlinear dimensionality reduction by locally linear embedding. Science, 2323–2326.

Rubin, R., Strayer, D., Rubin, E., 2011. Rubin's Pathology: Clinicopathologic Foundations of Medicine. Rubin's Pathology. Lippincott Williams & Wilkins. URL `http://books.google.com.tr/books?id=wb2TzY9AgJ0C`

Ruifrok, A., Johnston, D., 2001. Quantification of histochemical staining by color deconvolution. Analytical & Quantitative Cytology & Histology.

Sahiner, B., Chan, H. P., Petrick, N., Wei, D., Helvie, M. A., Adler, D. D., Goodsitt, M. M., 1996. Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images. IEEE Transactions On Medical Imaging.

Schwarz, M., Cowan, W., Beatty, J., 1987. An experimental comparison of RGB, YIQ, LAB, HSV, and opponent color models. ACM Transactions on Graphics 6, 123–158.

Sertel, O., Kong, J., Shimada, H., Catalyurek, U., Saltz, J., Gurcan, M. N., 2009. Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. Pattern Recognition 42, 1093–1103.

Shepard, R. N., 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. Psychometrika 27.

Siedlecki, W., Sklansky, J., Nov 1989. A note on genetic algorithms for large-scale feature selection. Pattern Recognition Letters 10, 335–347.

Society, A. C., January 2012. Colorectal cancer.
URL `http://www.cancer.org/Cancer/ColonandRectumCancer/DetailedGuide/colorectal-cancer-diagnosed`

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

Tenenbaum, J. B., de Silva, V., Langford, J. C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science.

UNSW Embryology website., 2009. Gastrointestinal tract, colon histology.
URL `http://php.med.unsw.edu.au/embryology/index.php?title=Colon_Histology_2009`

Vapnik, V. N., 1998. Statistical Learning Theory. Wiley.

Waheed, S., Moffitt, R. A., Chaudry, Q., Young, A. N., Wang, M. D., October 2007. Computer aided histopathological classification of cancer subtypes. In: Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on. pp. 503 – 508.

Wang, Y.-Y., Chang, S.-C., Wu, L.-W., Tsai, S.-T., Sun, Y. N., 2007. A color-based approach for automated segmentation in tumor tissue classification. In: EMBS 2007. 29th Annual International Conference of the IEEE. Engineering in Medicine and Biology Society, pp. 6576 – 6579.

Weyn, B., Wouwer, G. V. D., Daele, A. V., Scheunders, P., Dyck, D. V., Marck, E. V., Jacob, W., 1998. Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. Cytometry 33, 32–40.

World Health Organization, 2008. Cancer incidence, mortality and prevalence worldwide in 2008.
URL `http://globocan.iarc.fr/`

Wouwer, G. V. D., Weyn, B., Scheunders, P., Jacob, W., Marck, E. V., Dyck, D. V., 2000. Wavelets as chromatin texture descriptors for the automated identification of neoplastic nuclei. Journal of Microscopy 197, 25.

Yang, L., Meer, P., Foran, D., 2005. Unsupervised segmentation based on robust estimation and color active contour models. IEEE Transactions on Information Technology in Biomedicine 9 (3), 475–486.

# VITA

Devrim Önder was born in Ankara, Turkey, in April 8th, 1970. After having completed his high school education in Izmir High School of Science, he attended the Electrical and Electronics Engineering Department at Bilkent University, Ankara, Turkey. Devrim graduated from Bilkent University and acquired his Bachelors of Science degree in Electrical Engineering in June 1992. He acquired his Master of Science degree in July 1997 in the master of science program of Electrical and Electronics Engineering Department at Middle East Technical University, Ankara, Turkey.

Devrim worked as an R&D engineer in Image Processing Group Of TÜBİTAK BİLTEN Information Technologies Research Institute, Ankara, Turkey, between 1995 and 1997. Devrim worked as both Software Engineer and Software Team Leader in STM Defence Technologies Engineering Inc., Ankara, Turkey between 1999 and 2004. Between 2004 and 2008, Devrim worked as Software Engineer and Project Manager in Cabot Inc., Izmir, Turkey in several Digital Video Broadcasting (DVB) projects.

Since then, he has been working as a researcher in the fields of image processing, computer vision, statistical signal processing and software engineering.