CHALMERS

Classification of Oral Lesions from Histological Images

Master of Science Thesis in Biomedical Engineering

NOOSHIN JAFARI

Department of Signals and Systems
Division of Biomedical Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden, 2010
Report No. EX021/2010

Abstract

In this work, a Computer Aided Diagnosis (CADx) system for classification of oral cavity lesions from histological images based upon the image analysis and pattern recognition has been developed. The evaluation of oral conditions/diseases is primarily performed by a visual examination, followed by a pathological investigation of the suspicious area. However, this diagnostic criterion will almost certainly be subjective and dependent on the pathologist's knowledge, experience and interpretation. So, a computer based system may be used as a supportive tool for oral specialists to establish a more consistent and accurate diagnosis of oral cavity lesions.

From the extensive range of the lesions and abnormalities rising in the site of human oral cavity, we proposed to investigate two of the common and potentially precancerous lesions; Oral Submucous Fibrosis (OSF) and Oral Lichen Planus (OLP). The classification problem studied in this paper is considered as a two class problem; the Normal Oral Mucosa (NOM) tissues versus the OLP/OSF premalignant lesions and the investigated classifiers are Support Vector Machine (SVM) and *k*-Nearest Neighbors (*k*NN). We proposed to investigate the histogram-based properties of the tissue as discriminating features. Also, two color representation modalities (RGB and HSV color histogram systems) are used to evaluate their discriminative power for analysis of histological images.
Estimation of the classifier performance was done using Receiver Operating Characteristic based on the resubstitution, 5-fold cross validation and leave-one-out methods.

Relying only on the histogram-based properties of oral lesions, the overall classification accuracy was 83.7% (135/161) for juxta-epithelial ROIs with the sensitivity and specificity of 89% and 74%, respectively. Roughly the same accuracy (80%, 131/161) was achieved when the classifier was trained on sub-epithelial connective tissues. Employing the color histogram systems, the best results were achieved in HSV system (78% accuracy) using 5-fold CV and *k*NN classifier.

Acknowledgement

I am deeply thankful to my project supervisor, Dr. Artur Chodorowski, from Chalmers University of Technology, Göteborg, Sweden for his admiring suggestions and guidance in all matters relating to the preparation of this thesis work.

Thanks are also given to our Indian collaborators, Prof. Vinay Hazarey from Govt. Dental College and Hospital, Nagpur, India and Dr. Sunil Kothawar from Sharad Pawar Dental College and Hospital, Wardha, India for the contribution and making the histological data available for this project.

Table of Contents

List of Acronyms

HPV	-	Human Papilloma Virus
NOM	-	Normal Oral Mucosa
OLP	-	Oral Lichen Planus
OSF	-	Oral Submocous Fibrosis
ROI	-	Regions of Interest
kNN	-	k-Nearest Neighboring
SVM	-	Support Vector Machine
QP	-	Quadratic Programming
VC	-	Vapnik Chervonenkis
RBF	-	Radial Basis Function
CAD	-	Computer Aided Diagnosis
CV	-	Cross Validation
LOOCV	-	Leave One Out Cross Validation
HOCV	-	Hold Out Cross Validation
OED	-	Oral Epithelial Dysplasia
ROC	-	Receiver Operating Characteristics
NSV	-	Number of Support Vectors

1. Introduction

1.1. Medical Background

The incidence of cancer in the oral cavity (mouth, tongue, and pharynx) is dramatically increasing every year. Even though only five percent of all types of known cancers belong to the head and neck area, 30 percent of them are diagnosed in the oral cavity. In other words, oral cavity is the most frequent site of head and neck involved with the abnormalities affecting the oral mucosal membrane [AHNS, 1998; Paul et al., 2005].

The number of diagnosed oral malignancy varies from one country to another. India is one of the countries with the highest rate of reported oral cancers, possibly due to the excessive use of a specific type of smokeless tobacco and beetle nut. According to a typical statistic, estimated incidence of oral cancer in the United States is roughly 35,000 cases per year, with the mortality of 8000, approximately representing one death per hour. However, considering the worldwide statistics the predicament gets more significant since over 400,000 new cases are diagnosed with oral cancers each year [Kanwar, 2009; Brickley,1996], with the slightly higher prevalence reported among men (12%) in comparison with women (8%). Also, the reported incidence has been higher within the age groups above 45 [Kanwar, 2009].

The major predisposing factors for oral mucosal carcinoma are the excessive use of the tobacco products, heavy alcohol consumption and cigarette smoking. Beside life style factors, exposure to sunlight, being male, increased age, and being infected with human papilloma virus (HPV) can increase the risk for oral cancers, too. However, the main etiological factor of oral cancer is still remained as an enigma [NCI Committee, 2009].

The evaluation and diagnosis of oral mucosal lesions is primarily performed by a complete clinical examination. This should include a comprehensive physical examination by not only looking at the suspicious area of the oral cavity (visual examination), but also a palpation of the oral cavity exterior. Usually this is done for the entire head, neck and intraoral area in

order to check if the cancer is speared or not. Following the completion of clinical examination, sometimes, a CT scan or MRI or both is advised to rule in or out any other underlying tissue. This can also help to specify the scope of the lesion [Epstein, 2002]. In some pre-cancerous and asymmetrical cases, it is a challenge for physicians to establish a firm diagnosis, since the clinical appearance is not diagnostic alone [Epstein, 2002; Kanwar, 2009] for the oral lesions which may resemble clinically. For such doubtful cases, a biopsy - based on the histopathological assessment of the target area- is often advised to be sent to a pathologist. This would help to confirm or reject the clinical diagnosis. However establishing a definitive diagnosis is yet vague as there is a lack of objectivity on the histological interpretation. This makes the diagnosis outcome to be highly dependent on the pathologist who might fail to set up an accurate diagnosis [AHNS Committee, 1998; Epstein, 2002].

A vast majority of the precancerous oral lesions proceed to cancer or reach the advanced stages before they are diagnosed. In most cases, this situation is due to the late diagnosis of the abnormality, especially for the asymptomatic lesions, and might be caused by either a delayed report from the patients or by the clinicians/dentists who have failed to diagnose the premalignant lesion in its early developing stages [Epstein, 2002].

The potential premalignant lesions rising in the oral cavity area may show different symptoms. However, the most common one might be the painless, non-healing patches on the surface of the tong, inner cheek or mouth. This symptom is one of the key factors in remaining the early stage cancers and precancerous lesions unnoticed until they develop to the advanced levels. The patient may even disregard other signs such as color or tissue changes of the oral mucosal [Kanwar, 2009; AHNS Committee, 1998]. Two of the most common and premalignant disorders of the oral cavity are Oral Lichen Planus (OLP) and Oral Submucous Fibrosis (OSF). They arise from the Normal Oral Mucosa (NOM) tissues and have the potential to progress very rapidly.

Oral Lichen Planus (OLP) is a chronic inflammatory disease affecting the oral mucosal membrane (Figure 1.1). It still has no definite etiology and cure; however one may control it by medications. Usually OLPs have different clinical presentations but typically resemble at microscopic level. Although OLPs appear as white lace-like patches on the lining of the mouth and in many cases may be diagnosed clinically, they still represent six different clinical forms which might vary for each individual case. Histopathological diagnosis of OLP is based on the interpretation of the oral epithelial and underlying connective tissue. Concurrently, clinical and histological features of OLPs can establish a more accurate diagnosis; however both features have remained enigmas yet [MFMER, 1998; Rhodus et al., 2003].

Oral Submucous Fibrosis (OSF or OSMF) is a chronic enfeebling precancerous disorder affecting the oral cavity with highly potential risk of developing to cancer, in which the sub mucosal tissues are involved with an inflammation and progressive fibrosis bands (Figure 1.1) [Lountzis, 2009]. As the disease progresses, theses fibrosis bands become stiff, creating a palpable rigid surface on the oral mucosal. In advanced levels, the patient will eventually find difficulty in opening the mouth. The main etiology of OSF is not recognized yet. However, in Indian subcontinent, the habit of chewing tobacco products in combination with areca nut and betel leaf is understood as the main reason of OSF incidence especially among youths [Paul et al., 2005].

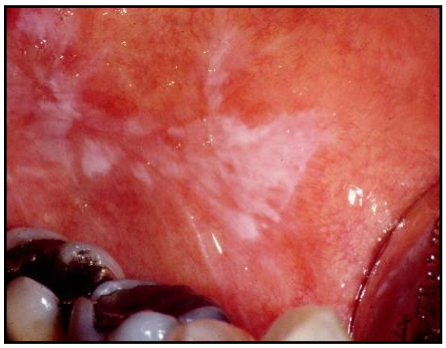
a.  b. 

Figure 1.1: Typical color images from a) Oral Lichen Planus (LOP) lesion, b) Oral Submucous Fibrosis (OSF) lesion

1.2. Motivation

Evaluation and investigation of oral mucosal lesions is primarily relying on the visual examination, followed by a histopathological evaluation of the suspicious area. Although this plays a critical role in the early detection of the premalignant lesions, the obtained diagnosis will almost certainly be subjective and dependent on the pathologist's knowledge, experience and interpretation. Lack of objectivity in identification of the premalignant oral lesions, may bring about high error estimates and less reliable diagnosis. So, the clinical and pathological evaluations cannot be investigative alone. In addition, as there is a high rate of malignant transformation in potentially precancerous oral lesions, it is so essential to identify the presence of such lesions before they reach the advanced levels. Consequently, there would be a necessity for developing a computer aided diagnosis system to assist the oral specialists to establish a more reliable diagnosis of the oral cavity lesions.

1.3. Goal

The list of different types of lesions or abnormalities, ranging from benign to cancerous, investigated within the area of human oral cavity is quite extensive. Many of these lesions can be diagnosed by the routine clinical examinations. However, some cases, such as Oral Lichen Planus (OLP) and Oral Submucous Fibrosis (OSF), are more elusive and challenging to be distinguished. In this paper the focus will be on the evaluation of these two common, potentially premalignant lesions. We propose to develop a Computer Aided Diagnosis (CADx) system based on investigation and classification of the OPL and OSF case samples from a viewpoint of Normal Oral Mucosa (NOM). Such a system may be used as a support tool to assist the clinicians to improve their daily diagnosis or as a computerized analysis tool to assist oral specialists.

1.4. Previous Work

To date, few studies have been carried out toward identification of oral conditions/diseases from histopathological findings. An attempt in the field was done by Abbey et al. [Abbey et al., 1995]. They investigated the presence of epithelial dysplasia on 120 oral biopsies examined by six board-certified oral pathologists. The diagnoses were based on the histopathologic information of oral epithelial dysplasia (OED) - no clinical findings were submitted. They reported an agreement of 50.5% between oral pathologists' diagnoses with the original sign-out diagnosis and 81.5% of agreement regarding the presence or absence of epithelial dysplasia. Adding the clinical information to the current experiment [Abbey et al., 1997] even decreased the accuracy and consistency of the diagnoses made by oral pathologists. Another study was carried out by Zerdoner [Zernoder, 2003] on Ljubljana classification which is a system for grading of laryngeal lesions. He proposed to evaluate the applicability of the system to classify the oral cavity lesions based on their histological changes in the epithelium. The conclusion of his study on 135 oral lesion biopsies turned out that Ljubljana can be used as a reliable grading system for classifying the oral epithelial lesions. Paul and his coworkers [Paul et al., 2005] designed a CAD system based on wavelet artificial neural network for identification of precancerous oral tissue (OSF) from normal stages, using transmission electron micrographic images of collagen fibers. They reported that the properly classified cases in their proposed technique have always been greater than 50%.

1.5. Structure of Thesis Report

The whole thesis report is organized in 4 chapters, appendices and finally references. Chapter 1 is the introduction part of the thesis. It gives a brief medical background of the oral cavity lesions, the thesis motivation and the thesis goal. Chapter 2 starts with the depiction of the material, followed by the description of the features investigated in this project. Next, the principle of the two classification algorithms, including the *k*NN and SVM, are concisely explained. Finally, different alternative approaches for estimation of the

classifier performance are briefly discussed in this chapter. In chapter 3 we present the experimental results obtained from the implementations in Matlab together with the general discussions. Finally, Chapter 4 follows the conclusion and proposed future work.

2. Materials & Methods

2.1. Materials

The biopsy specimens analyzed in this study have been taken from the patients who were referred to the departments of oral pathology at Govt and Sharad Pawar dental colleges and hospitals in Wardha and Nagpour, India. All biopsy samples have gone through a routine light microscopic examination subsequent to a process of staining by aqueous haematoxylin and eosin. Next, in order to avoid the human bias and also verify the previous diagnosis made by clinicians/physicians, four oral histopathologists have performed a histopathological evaluation process on available samples. The histopathological images have been recorded with typical resolution of 10x by using Olympus BX51 Research Microscope with attached Olympus DP71 Camera and acquired by means of 'Cell^D' Image Analysis Software.

Totally, 36 numbers of histopathological oral mucosal images were investigated in the present work. The entire cases were categorized into three individual classes, namely, Normal Oral Mucosa (NOM), Oral Lichen Planus (OLP) and Oral Submucous Fibrosis (OSF). Table 2.1 summarizes the number of existing samples per class and their corresponding number of extracted ROIs. The classification problem was regarded as a 2-class problem, considering healthy tissues (NOM) against the pre-malignant lesions (OSF/OLP), since this case is amongst the most challenging diagnosis for oral specialists to distinguish. Figure 2.1 shows some typical histopathological images representing three oral cavity conditions investigated in this study.

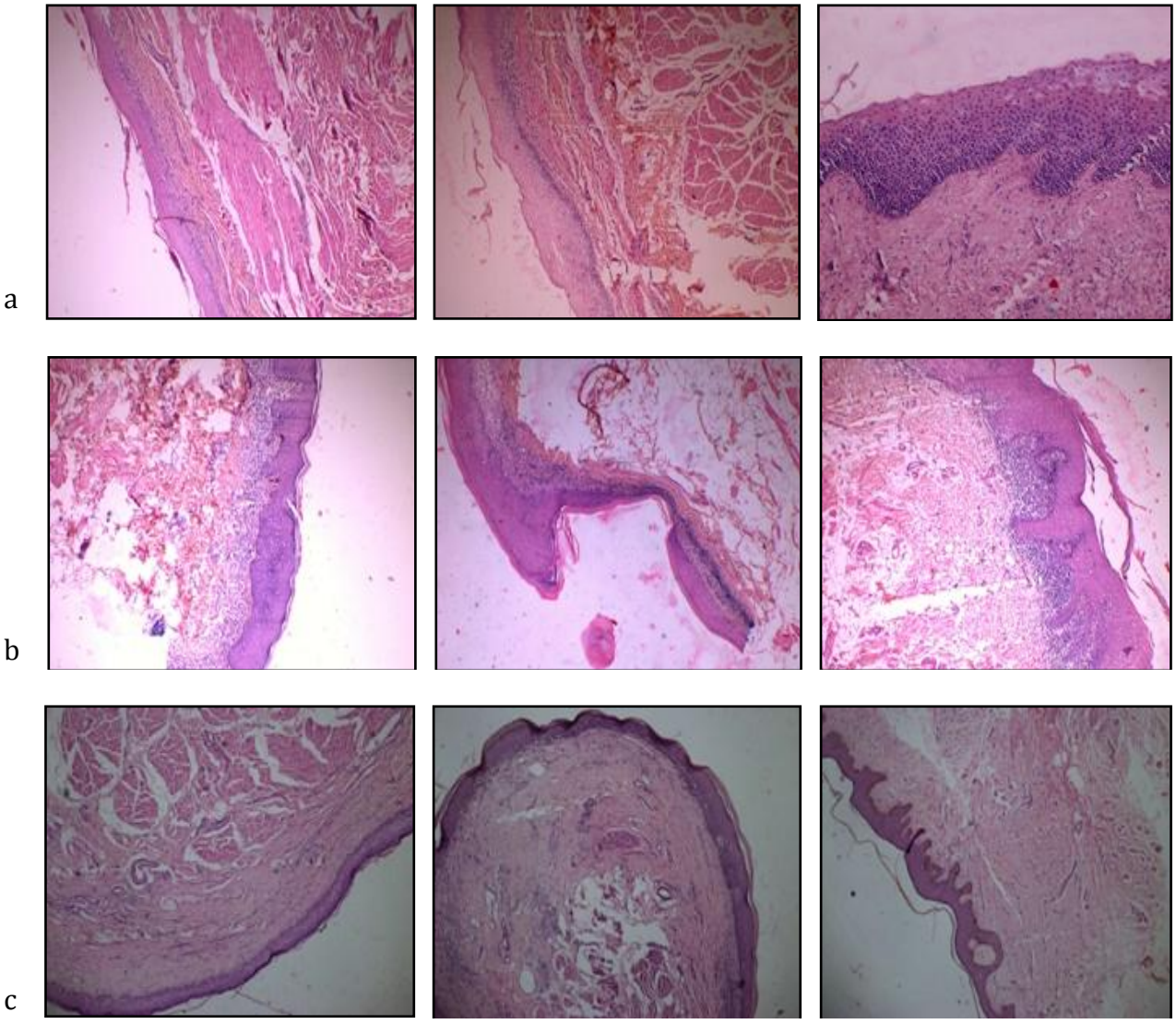


Figure 2.1: Typical histological images taken from oral cavity conditions representing three different subtypes; a) Normal Oral Mucosa (NOM), b) Oral Lichen Planus (OLP) and c) Oral Submucous Fibrosis (OSF)

2.2. Methods

In this section we present the image pre-processing and image analysis steps, which have been followed in the current work, toward the classification of histological oral images. The overall procedure is demonstrated in block diagram in Figure 2.2.

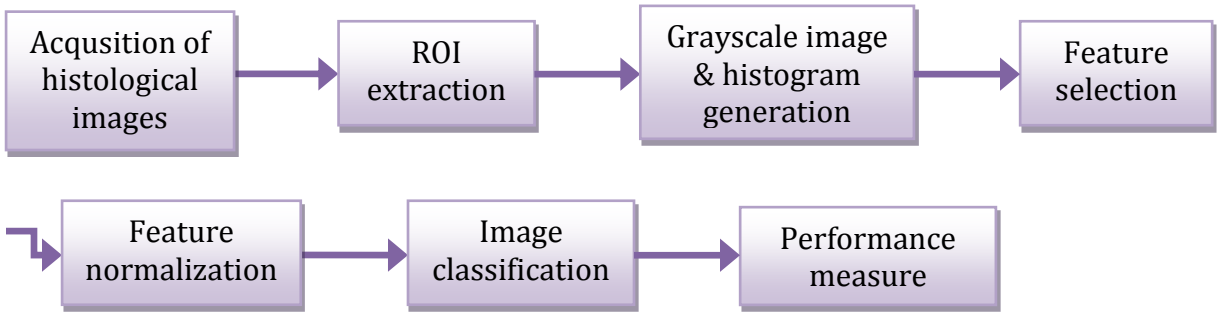


Figure 2.2: Stages of histological oral image pre-processing and analysis

2.2.1. Regions of Interest

In this work, experienced oral histopathologists have manually marked out the Regions of Interest (ROIs) from the histological images of oral epithelial tissues. The first extracted ROIs regarded as C1 are placed in the area beneath the epithelial exterior called juxta-epithelial connective tissue, which are depicted by the yellow rectangles in Figure 2.3 The other regions of interest, C2, have been outlined from the epithelial interior area (sub-epithelial connective tissue), which are demonstrated by the blue squares in Figure 2.3 ROIs have been extracted from both lesion areas (OLP/OSF samples) and healthy tissues (NOM samples). The areas which are regarded as artifact do not represent any identical properties with the two other regions. Consequently, they should be excluded from the training database. Because of the low number of currently available samples used in this experiment, we have extracted many ROIs regarded as C1 and C2. This way, we could increase our training set from a total number of 36 samples to 161 for both juxta-epithelial and sub-epithelial regions (Table 2.1). Some examples of the C1 and C2 regions, extracted from each individual study case, are presented in Figure 2.4. In this work, we have mostly conducted the classification experiments on juxta-epithelial (C1) regions.

Table 2.1: Total number of study cases and the corresponding number of ROIs per class

Histological diagnosis	# of cases per class	# of ROIs per class	
		Juxta-epithelial area	Sub-epithelial area
Normal Oral Mucosa (NOM)	14	61	67
Oral Submucous Fibrosis (OSF)	12	53	42
Oral Lichen Planus (OLP)	10	47	52
Total # of cases	36	161	161

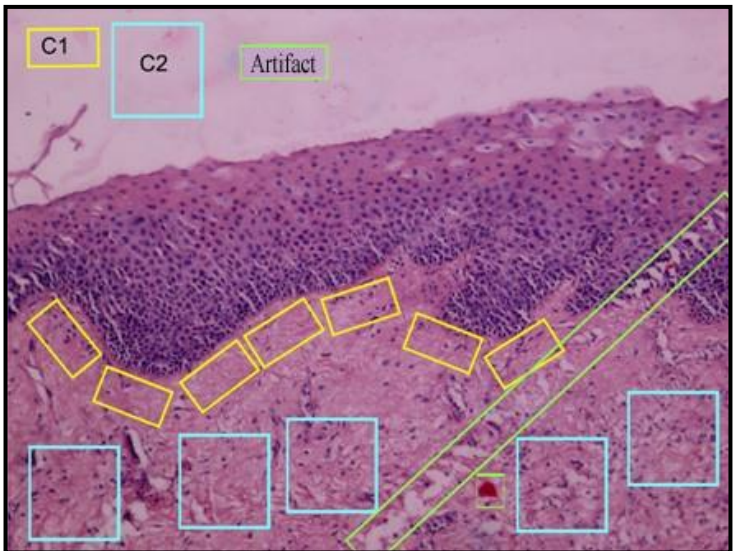


Figure 2.3: An example of oral histological image with the marked out regions of interest: 1) yellow rectangles, C1, extracted from juxta-epithelial tissue, 2) blue squares, C2, extracted from sub-epithelial tissue. The regions regarded as artifact does not represent any discriminative property.

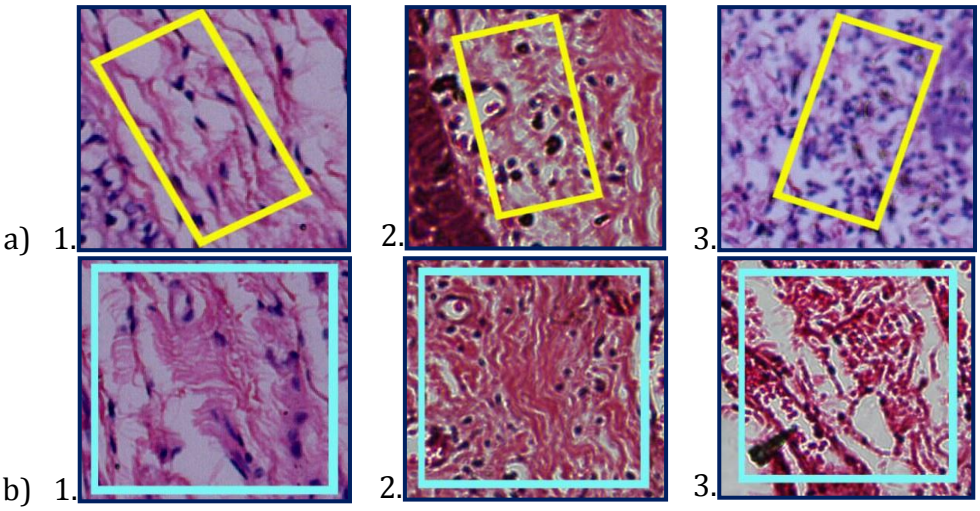


Figure 2.4: Typical histological images representing the NOM, OSF and OLP cases, respectively, extracted from a) C1 regions of interest, b) C2 regions of interest

2.2.2. Histogram Generation

In this study we are mainly concerned with the task of analyzing the histogram-based properties of the pre-cancerous oral lesions with the emphasis on the variations in the shape of the image histogram. The aim is to produce features which are relatively robust to small changes in the image.

The first-order color histogram is a graphical representation of the frequency distribution of the image pixels. It is essentially a statistical probability distribution of the intensity (or gray level) values in an image versus the number of pixels. It contains robust and efficient information about the nature of an image which probably contain discriminatory information to be used in the field of image classification. Although histograms are relatively invariant to small changes such as rotations, translations and scale variations which are practical characteristics in image classification, they are lacking the spatial information of the object. This means that spatial relations between the pixels of an image are lost. So, this may result that many different images with different object contents possess similar spatial distributions. The histogram (distribution) value, *P(i)*, of an image can be defined as [Sergyan, 2008; Liu and Wang, 2009]:

$$P(i) = \frac{N(i)}{M} \; for \; i = 1,2, \dots, N_g \tag{1}$$

where *N(i)* is the number of pixels at gray level *i*, considering to have N_g number of distinct gray levels ranging from 1 to 256 in the quantized image, and *M* denotes the total number of pixels in the image. We have normalized the histogram value, P_i, so that the summation of histogram values for any specific probability distribution will be equal to 1.

In order to reduce the image dimensionality and accordingly simplify the calculations, we have primarily performed an rbg to gray scale conversion. We used Matlab rgb2gray [MATLAB, 2004] routine on the target areas. The image gray scale values are then transformed to 256 histogram bins. The histogram generation procedure is demonstrated in Figure 2.5.

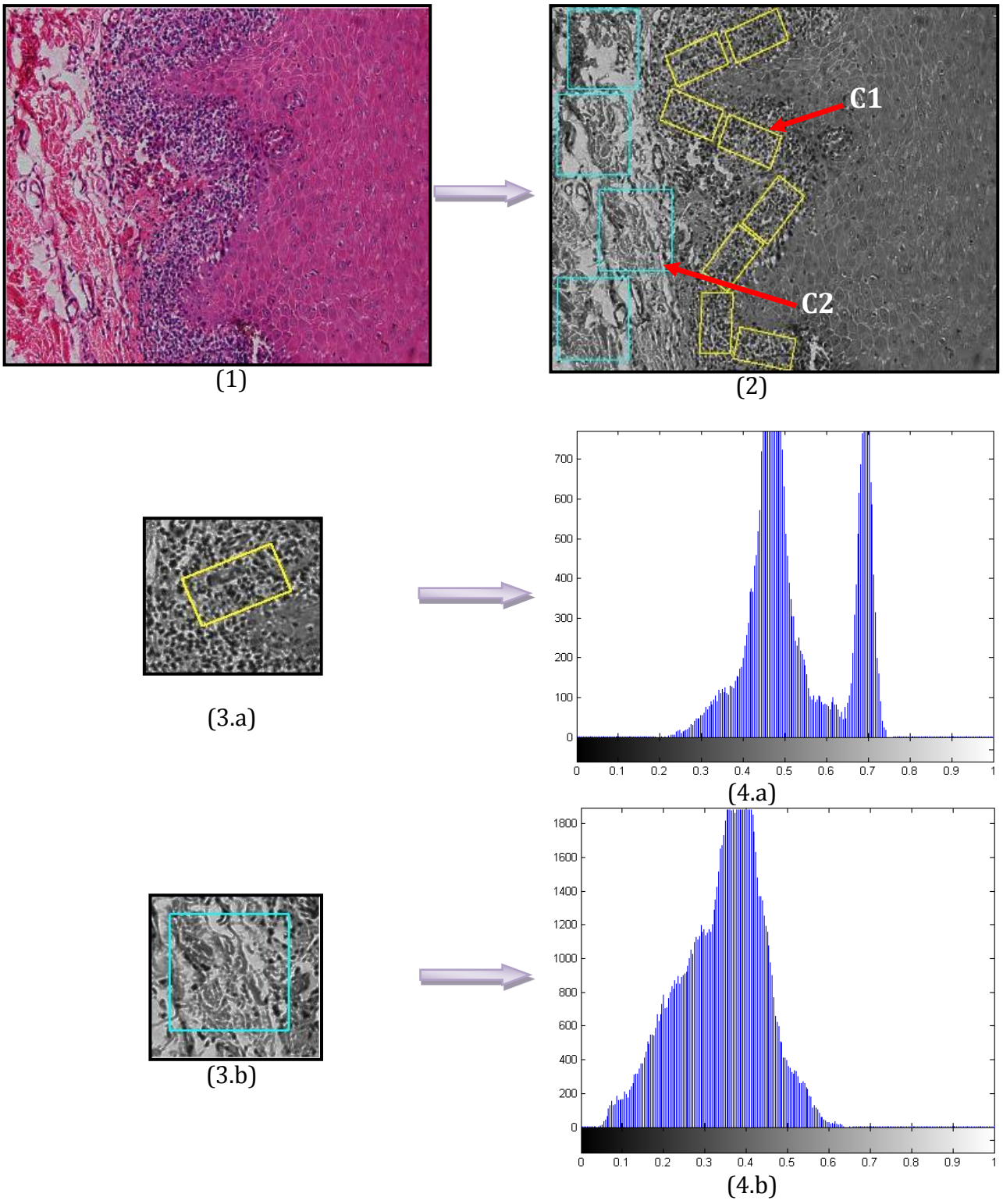


Figure 2.5: Illustration of proposed procedure toward histogram generation. (1) A typical histological image. (2) The corresponding gray scale image. (3) Typical ROIs extracted from: a) juxta-epithelial area (C1), b) sub-epithelial area (C2). (4) Histogram generation.

2.3. Features

2.3.1. First-Order Histogram-Based Features

The first-order histogram of an image provides us practical information to be used for image texture analysis. Primarily, Swain et al. [Swain & Ballard, 1990] proposed the histogram characteristics to be used as one of the main feature descriptors in image processing. In this work, we have taken cumulative dark introduced in [Chodorowski, 2009] and a number of common histogram based features including mean, variance, entropy, energy, skewness, and kurtosis as discriminative features. It should be noted that not all of the features contribute equally to image classification. In the followings (section 2.4) some techniques for weighting or restricting the feature set are represented [Paul et al., 2005]. Below are the definitions of the extracted features:

The ***mean*** is a measure of brightness in an image i.e. a very dark image implies a low mean value while a bright image implies a high mean value. It calculates the average of intensity level distribution of an image and is defined as follows:

$$\mu = \sum_{i=1}^{N_g} iP(i) \tag{2}$$

The ***variance*** (also equals the square of the standard deviation) tells us something about the uncertainty or dispersion of the intensity levels in an image. In other words, variance is a measure of the image contrast i.e. an image with high contrast implies a high variance value whereas an image with low contrast implies a low variance value.

$$Variance = \sum_{i=1}^{N_g} (i-\mu)^2 P(i) \tag{3}$$

The ***entropy*** quantifies the amount of the disorder or randomness of gray scale values in an image. A uniform image with more different intensity levels has higher entropy value than a simple image. We can define the entropy as follows:

$$Entropy = -\sum_{i=1}^{N_g} P(i) \cdot \log_2[P(i)] \tag{4}$$

The ***Skewness*** is a measure of asymmetry about the shape of the frequency or intensity level distribution, and is given by:

$$Skewness = \sum_{i=1}^{N_g} (\mu_i - \bar{\mu})^3 / (N-1)s^3 \tag{5}$$

where s is the standard deviation (or the square root of the variance) and N denotes the number of data points. A distribution will be "positively skewed" if the tail of its histogram extends out to the right, and "negatively charged" if the tail of its histogram extends out to the left. The skewness can also be defined by the following formula [Pearson, 1895]:

$$Skewness' = (\mu - mode)/var^2 \tag{6}$$

where mode denotes the peak value in the histogram.

The ***kurtosis*** measures how flat or peaked the top of the data distribution is, compared to the normal (or Gaussian) distribution. A positive kurtosis denotes a flat-topped data distribution while a negative kurtosis refers to a peaked data distribution. It can be defined as:

$$Kurtosis = \sum_{i=1}^{N_g} (\mu_i - \mu)^4 / (N-1)s^4 \tag{7}$$

The ***energy*** or ***Angular Second Moment (ASM)*** quantifies the homogeneity within the texture of an image. The homogeneity of an image will decrease as the number of intensity levels increases which corresponds to a smaller energy value. For example, the energy measure of a uniform image will be large as the energy of pixels is concentrated in a few number of intensity levels.

$$Energy = \sum_{i=!}^{N_g} [P(i)]^2 \tag{8}$$

The ***cumulative dark*** [Chodorowski, 2009] represents the number of image pixels which their relevant values in the gray scale histogram drops below the center value.

$$Cumulative\ Dark = \sum_{i=1}^{N_g/_2} P(i) \tag{9}$$

2.3.2. Color Histograms

A more simple approach to image classification based on the image color histogram is to directly use the raw histogram bins as input features to the classifier. However, the major drawback of this approach in image classification would be the high dimensionality (or the number of bins) of input vector. In this work, from different choices of color systems, we investigated images in the red-green-blue (RGB) and hue-saturation-value (HSV) color spaces and generated their one-dimensional and three-dimensional histograms, considering 16 bins per color channel [Chapelle, 1999]. This results in feature vectors of 48 (16·3) and 4096 (16^3) feature indices, respectively.

The RGB color space is obtained by separating the three primary colors of light (red, green and blue) into discrete arrays. The frequency of discrete color components is then acquired by counting the number of times that each color occurs in the image array. Each color

component, with 8 bits of unsigned numbers, is represented in 256 different values, which brings about 16.7 million colors.

The HSV color space also represents the color into three components. Hue, the first component, determines the angular position (wavelength) of the color. The second component, saturation, has the range of 0 to 100 and is an indication of the color depth or purity in the image. The saturation value of 0 represents the grey color and as it increases towards 100, the gray shade decreases in the color space and the color tends toward a primary color. The third component, value (also known as intensity), describes the brightness of the color space. The hue and saturation components (color information) are separated from intensity component (luminance). Compared to RGB system, HSV space remains roughly unchanged under illumination changes. In the sense of human perception of colors, the HSV color system is more frequently used than RGB space [Surak, 2002].

2.4. Feature Selection

One of the main tasks in a pattern recognition problem for data mining, before designing the classifier, is feature selection. The idea of feature selection is to identify a subset of features of potential interest, relevant to a particular application, which represents the best performance under some classification systems [Vandewalle et al., 2003]. This procedure reduces the feature space dimensionality by choosing the relevant features from a large set of possibly redundant or irrelevant candidate features based on some criterions. Some advantages of applying the feature selection techniques in the machine learning problems can be stated as constructing more robust CAD tools, achieving better classification accuracy, improving generalization performance, and speeding up the process of data mining. In addition, it can compensate the effect of finite sample size, especially in the medical CAD systems, by eliminating the useless features and reducing the size of the structure [Miyamoto et al., 2003; Yuan et al., 1999]. We performed a supervised feature selection, as the entire samples used in this study are pre-labeled.

The problem of feature subset selection can be formulated as follows. Let's denote the size of entire feature set by N, the original feature vector by $X_N = \{x_1, x_2, \dots, x_N\}$, and the optimized feature subset vector of size M by $Y_M = \{\{x_1(i), x_2(i), \dots, x_M\} \mid i = 1,2, \dots, N\}$. Here, the task involves finding out a searching strategy which determines a good subset of features. In statistical subset selection, a "good" subset usually means the subset which optimizes the objective function *J(y)* and results in the highest classification accuracy [Kim et al., 2006].

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ \vdots \\ x_N \end{bmatrix} \longrightarrow \begin{bmatrix} x_1(i) \\ x_2(i) \\ \vdots \\ x_M(i) \end{bmatrix} \tag{10}$$

$$Y_M = argmax[J\{x_i | i = 1, \dots, M\}] \tag{11}$$

To find out the best feature subset of size M, one requires checking up

$$\left\langle \begin{matrix} N \\ M \end{matrix} \right\rangle = \frac{N!}{M!\,(N-M)!} \tag{12}$$

number of examinations. Totally, there are 2^N number of possible subsets from N features [Jain & Zongker, 1997].

Depending on whether or not the feature subset selection is reliant on the classifier performance, the existing approaches can be separated into filter and wrapper algorithms. In filter approach the effect of any prediction method on selecting the features in not taken into account. This method is on the basis of relevance or discrimination power of the features and separability of clusters. It filters the irrelevant features based on the feature selection criterion. On the other hand, in wrapper approach the predicted accuracy directly relies on the usefulness and discrimination power of the features. A certain classifier is employed to select a subset of features by usually performing an exhaustive search for all

possible feature combinations. Usually filter approach outperforms the wrapper approach [Liu et al., 2005; Yuan et al., 1999].

Tow general wrapper approaches to feature subset selection are sequential forward selection and sequential backward elimination methods. In sequential forward selection, we start with an empty feature vector and continue with adding up the features one by one. At each step, the feature which represents the best accuracy is selected. This is continued until any additional feature does not lead to significant reduction in the classification error. In sequential backward elimination, we start with the entire set of features and at each step remove the feature which leads to the most decrease in the classification accuracy. This is continued until no further removal improves the accuracy significantly [Sewell, 2007].

In this work, we performed a sequential forward selection on the histogram-based features. All possible combinations of features were evaluated through an exhaustive search. Out of the available combination of features, the subset consisting of the entropy, mean, skewness, and variance produced the best classifier's performance. The 5-fold cross validation error was chosen as error estimation criterion.

2.4.1. Feature Normalization

The normalization of the entire feature vector was performed in order to set the variance and mean of all training samples, corresponding to the i^{th} feature attribute, to one and zero, respectively. This results that all attributes in different numeric ranges weight relatively the same. The advantage is that even the large attribute values will, therefore, place in a finite numeric range and will not dominate the ones in smaller ranges [Chih-Wei, 2009]. Normalization was performed according to the following formula:

$$\tilde{X}_i = (x_i - \mu_i)/s_i \quad i = 1,2,\dots,m \tag{13}$$

where $\tilde{X}_i$ is the normalized feature attribute, μ_i and s_i denote the mean value and standard deviation of i^{th} feature value, respectively, and m is the total number of features.

2.5. Classifiers

Although the selected feature sets have the greatest influence on the result of the classification performance, the choice of classifier can also play a critical role in the final result. Currently, there are over 200 different choices of classifier types [Axell, 1976] including parametric and non-parametric classifiers for supervised classification. Usually, depending on the classification problem, a certain classifier might be chosen. In the current work, the data was not well-separated. Accordingly, we selected the *k*-nearest neighbor (*k*-NN) as non-parametric and support vector machine (SVM) as parametric classifier.

2.5.1. *k*-Nearest Neighbor

The *k*-Nearest Neighbor (*k*NN) decision rule is one of the commonly used classification tools in pattern recognition problems. It is based on the distance between the training data points from the feature vector $\mathrm{x} = [\mathrm{x}_1\ \mathrm{x}_2 \ldots \mathrm{x}_\mathrm{d}]$ and a particular unknown data object (which is to be classified) $\mathrm{x_i} \in \mathrm{R^d}$, $i = 1,2, \ldots, k$, in a d-dimensional feature space R^d. In *k*NN algorithm an unseen data is classified by election of its neighbors. Here, the task is to define the class label of test data by means of a particular metric which is usually the Euclidean distance defined by:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (x_{r,i} - x_{r,j})^2} \tag{14}$$

where x_i and x_j are two sample points in the feature space.

The performance of *k*NN classifier is to some extent determined by the selection of an appropriate *k* value (the number of nearest neighbors to sample test data, x_o) which might be challenging. The majority vote defines the true class label of x_o. Normally larger value of *k* leads to a system with more immunity to noise. Figure 2.6 gives an example that depicts the choice of *k* parameter on prediction of class label of query point (green circle). To find the true class for an entry test data the classifier calculates all Euclidean distances from this

test data and all pre-labeled training data. The distances are then arranged in an ascending order matrix. The shortest distance represents the right class to which the unseen data belongs to [Song et al., 2007].

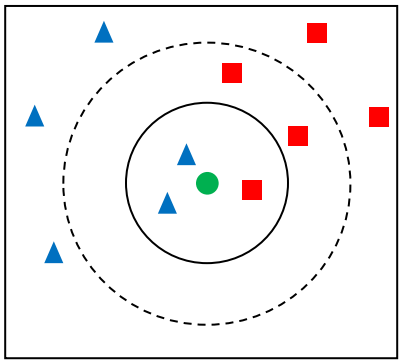


Figure.2.6: Example of k-nearest neighbor (kNN) classifier; choice of k parameter to decide which classes the query object (green circle) belongs to. The object will be classified as a rectangle if k=3, but as a rectangle if k=5.

The routine for the *k*NN classifier trailed in this project is as follows (assuming an *M*-dimensional training category and assuming to have one test sample to be classified at a time):

1. Calculating the Euclidean distance between the test data and the entire set of training data
2. Sorting the obtained Euclidean distances in an ascending order in a distance matrix
3. Determining to which class the *k* nearest distances are associated to (counting the number of votes)
4. Classifying test data to the class that was counted the most in step 3

One of the advantages of the *k*NN classifier is being compatible for multi-modal classes as its classification decision is based on a small neighborhood of similar objects. So, even if the target set is multi-modal (i.e. contains objects whose independent variables have unlike characteristics for unlike divisions), it can still classify with high accuracy. The main drawback of this method is being computationally slow and complex as it calculates the

distances between the test data and all training data. In applications with large number of training data this method can be rather slow.

2.5.2. Support Vector Machines

The Support Vector Machine (SVM) is a commonly used technique for statistical classification of data, introduced by Boser, Guyon and Vapnik in 1992 [Cortes & Vapnik, 1995]. It parameterizes the distributions. The classification problem, for an M-dimensional feature space, involves a set of training points $(x_i, y_i), y\epsilon\{-1, +1\}, x\epsilon R^m, i = 1,2, \dots, m$ and where each data instance x_i contains a target value (class label), y_i, and several attributes (features) [Hsu et al., 2009]. The data point x_i belongs to either two of the classes depending on its attributes. The main objective of the binary SVM algorithm is to construct a hyper plane in such way that data points with the same target values place on the same side of the hyper plane. A good generalization performance (the ability to accurately classify an unseen data) is achieved within the maximum separation margin between two classes. The optimal separating hyper plane (optimal decision function, $f(x)$) is defined by:

$$f(x) = \text{sgn}\left(\sum_{\text{i=1}}^{\text{m}} \alpha_i y_i \, k(x_i, x) + b\right) \tag{15}$$

where

$$\text{K}(\text{x}_\text{i}, \text{x}_\text{j}) \equiv \emptyset(\text{x}_\text{i})^\text{T} \emptyset(\text{x}_\text{j}) \tag{16}$$

is a symmetric positive function called kernel. b is the bias or threshold parameter and α_i are Lagrange multipliers, obtained from the following quadratic programming (QP) optimization problem:

$$Q(\alpha) = \sum_{i=1}^{m} \alpha_i - {}^{1}/_{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (17)$$

and must be maximized under the following constraints:

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \; and \; 0 \leq \alpha_i \leq C \; for \; i = 1,2, \dots, m \quad (18)$$

where $C > 0$ is the SVM-regularization or penalty parameter of the error term, specified by user. To achieve a better generalization performance in a SVM classification problem, C parameter shall be chosen appropriately. The amount of overlap between classes can be controlled by this term. Although there is a trade-off between the classification error and the complexity of decision rule, one may control this by changing the value of C. Large values of C will result in a complex boundary surface that is too fit to training data but may not to unseen data. Contrary, low values of C will result in a simple boundary surface and a high classification error [Cortes & Vapnik, 1995]. Another challenge, when designing a SVM classifier, is to find an appropriate kernel function that has a critical effect on SVM performance. Some of the most commonly used kernel functions are given in Table 2.2. Usually, Gaussian or polynomial kernels are considered as default options [Vapnik, 1999].

Table 2.2: Kernel Functions

Kernel type	**Function**
Polynomial	$(x.y + 1)^d$
Gaussian	$exp(-\|\|x - y\|\|^2/2\sigma^2)$
Laplace	$exp(-\rho\|\|x - y\|\|_l)\,, \rho > 0$
Sigmoid neural network	$tanh(\rho x.y - \sigma)$

Here σ, d and Υ are input kernel parameters to SVM training process. For each particular application, the kernel parameters must be tuned to fulfill the sufficient classification performance [Jakkula V].

The most frequently used kernel functions are Gaussian Radial Basis Function (RBF) and Polynomial kernel. The Polynomial kernel is usually used for non-linear modeling e.g. in SVM kernel machines which produce the maximum margin by using all mapped features in the polynomial mapped feature space. As there is still no certain way to find the best kernel type, we randomly tried a few kernels with varying their kernel parameters e.g. kernel width (σ) in Gaussian, polynomial degree (d) in polynomial kernel and ρ (offset) in sigmoid kernel.

Producing an optimal decision function, which is to minimize the classification error by accurately classifying the unseen data, is based on the theory of structural risk minimization (SRM) principle. Following this principle, minimization of expected risk $R(\alpha)$ of unseen data involves minimization of both empirical risk $R_{emp}(\alpha)$ and Vapnik-Chervonenkis (VC) dimension h (terms 1 and 2 in eq. 17). Expected risk $R(\alpha)$is defined as [Vapnik, 1999; Axelberg et al., 2007]:

$$R(\alpha) = \int \big(y - f(x,\alpha)\big)^2 dp(x,y) \;\; for\; \alpha \epsilon \Lambda \tag{19}$$

where y denotes the output vector associated for an input example x, $P(x, y)$ is a fixed but unknown distribution function , and $f(x, \alpha)$ are set of functions implemented by learning machine with respect to $\alpha \epsilon \Lambda$. The aim is to find the function $f(x, \alpha)$, in such way that $R(\alpha)$ will be minimized.

The Empirical risk and the SRM principle respectively are defined by:

$$R_{emp}(\alpha) = -\frac{1}{N}\sum_{i=1}^{l}\big(y_i - f(x,\alpha)\big)^2 \tag{20}$$

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(\ln((2l/h)+1) - \ln(\eta/4)}{l}} \; for \; 0 \leq \eta \leq 1 \qquad (21)$$

where N is the number of training examples, η denotes a confidence measure and h denotes VC-dimension.

2.6. Estimation of Classifier Performance

2.6.2. Overall Classification Rate

Following the selection of an informative subset of features and designing an efficient classifier, estimation of performance criteria is the next step involved in a pattern recognition problem. It is an essential but quite difficult task to predict the classifier's performance under the constraints of the available finite dataset and unknown probability distribution of the dataset [Sahiner et al., 2007; Fukunaga & Hayes, 1989 (b)]. The finite dataset itself may lead to biased estimates (difference in error estimations).

The classifier performance is primarily based on the relationship between the number of available samples and the classifier dimension (number of features). It is beneficial to construct the classifier with the larger and more representative samples. However, in a practical pattern recognition problem, particularly in the development of a medical CAD system, the sample size is limited. Therefore, it is the designer's task to decide on the number of samples to be used for training the classifier and for testing its performance. The datasets should be independent but representing the same probability distribution. For example, large number of training samples together with a small amount of testing samples may result in a reliable classifier but unreliable performance estimation [Sahiner et al., 2007; Chan et al., 2004; Estimation of Classifier Performance_1]. In contrast, a large dimension of feature space often has a disadvantageous effect on the performance of the classifier. This may be explained by potentially correlated and less diagnostic features [Mazurowski et al., 2007].

The error rate estimation is a practical way of representing the classifier performance. It is defined as the mean probability of the counted number of misclassified unknown samples $\hat{\tau}_i$ (the test samples with incorrectly predicted class labels) over the entireN_i test samples drawn from the probability distribution D [Fukunaga & Hayes, 1989 & 1990].

$$\hat{\varepsilon}_i = \frac{\hat{\tau}_i}{N_i} \tag{22}$$

Total probability of error rate can be defined as follows:

$$\hat{\varepsilon} = \sum_{i=1}^{m} P_i \frac{\hat{\tau}_i}{N_i} \tag{23}$$

where $i = 1,2, \dots, m$ denotes the label of observations and P_i denotes the priori probability of distribution D_i. Here we assumed that the designing and testing samples are independent. Thereby, the estimated error rate $\hat{\varepsilon}$ is unbiased and its expected value will be $E\{\hat{\varepsilon}\} = \varepsilon$. The predictive accuracy $\hat{A}$ is then defined as the percentage of the corrected classified instances, likewise:

$$A = 1 - \hat{\varepsilon} \tag{24}$$

To achieve a robust and consistent estimation of the classifier performance, the variance of the true error rate is provided, and is given by:

$$Var\{\hat{\epsilon}\} = \sum_{i=1}^{N} P_i^2 \frac{\epsilon_i(1-\epsilon_i)}{N_i} \tag{25}$$

One of the practical limitations associated with the performance prediction of the classifier is the effect of sample size on the error evaluations. It is a challenge to decide on the number of samples which should be used for designing and testing the classifier. This issue

is more dominant when constructing the medical CAD systems, since the sample size is relatively small. Depending on how the data divisions are made, there are different alternative approaches to performance prediction. They are called resampling techniques which are divided into resubstitution and three types of cross validation methods, namely, the hold-out, *k*-fold and leave-one-out cross validation. The effect of dependent training and test sets is identified in the resubstitution and leave-one-out methods, while in the hold-out method the training and test sets are completely separated [Fukunaga & Hayes, 1989 (a); Sahiner et al., 2007; Varma & Simon, 2006].

Resubstitution method: Testing the classifier model on the original design samples which have already been used for training the classifier gives the resubstitution error rate. The estimate of the error rate is optimistically very low (usually zero), since it is basically made from the design samples. Therefore, the evaluations do not indicate the performance on the unseen data and represent only some knowledge regarding the performance of the used algorithm. For small datasets this method may consequent to poor generalization ability, while for large datasets it shows good results.

Hold-out cross validation (HOCV) method: is the simplest type of cross validation method in which the entire data set is split into two disjoint training and testing groups. Usually, two third of the data is used for training and the remainder is used for testing. To make the estimates more consistent and reliable, different sub samples are randomly selected and used for training and testing during some repeated iterations. Different error rates will be made from different divisions of data set. In the other words, the error estimates are so dependent on the data partitioning and thus the variance is relatively high.

k-fold cross-validation method: is an improved type of the hold-out method in the sense that the data points are used more efficiently. The data set is split into *k* subsets of approximately equal size. Each time, one subset is used as testing and the remaining (*k*-1) samples are merged to create the training set. This process is repeated *k* times. Then the error estimates are averaged across all *k* trials to yield the *k*-fold cross validation error rate.

According to the theoretical evidences, the best choices for k are 5 and 10 to get good results. However the variance will reduce as the number of k increases. In medical applications where the number of database is usually too low, the main drawback of k-fold cross validation would be the incompetent use of data. It is because each time only $1/k$ of all available data is devoted for training which is really few for the classifier design.

Leave-one-out cross validation (LOOCV) method: is a specific type of k-fold cross validation in which k is replaced to n (the total number of available samples). It means (n-1) samples design the classifier model and the prediction process is made for the left-over point. This process is repeated n times until all points are used for testing. This way, we will make the best use of our data points. The misclassified test points are then counted from the individual iterations and averaged to evaluate the overall error rate. The LOO cross validation method gives good evaluations of the model. However, it is usually computationally extensive when the number of database examples is high. The LOO estimate is not dependent on the classifier model or distribution of data, but its variance is quite large.

2.6.2. ROC Analysis

Receiver Operating Characteristics (ROC) curve has been introduced in machine learning as a robust way of estimating the performance of a classifier. ROC curves provide us the possibility to tune the performance of our trained classifier based on its corresponding trade-offs. In some applications such as medical diagnosis, it is more realistic to use ROC analysis than other commonly used measures for evaluating a classifier performance such as accuracy and error. This is because in such situations, ROC analysis decomposes the performance of a classifier into true positive (TP) and false positive (FP) fractions.

Considering a binary classification, for each instance there are four possible outcomes in general: true positive (TP) indicates the number of negative instances (or abnormalities in medical applications) which are correctly predicted, false positive (FP) indicates the

number of positive instances which are incorrectly predicted, false negative (FN) is the number of negative instances that classifier has wrongly predicted and true negative (TN) denotes the number of correctly classified positive cases. This situation, represented in Figure 2.7, is known as contingency table or confusion matrix [Qin, 2005].

		Classifier Predicted labels	
		Positive predictions (NOM)	Negative predictions (OLP/OSF)
Correct predicted labels	Positive cases (NOM)	True Positives (TP)	False Negatives (FN)
	Negative cases (OLP/OSF)	False Positives (FP)	True Negatives (TN)

Figure 2.7: Format of a confusion matrix representing the four possible classifications of data instances from a binary classification

The classical performance metric derived from the above 2 class confusion matrix is a function of true positive rate (TPr) versus the false positive rate (FPr). However, there are several other alternatives derived as performance metrics. In medical applications, two of the commonly used performance measures are known as sensitivity and specificity defied as:

$$sensitivity = \frac{TP}{TP + FN} \tag{26}$$

$$specificity = \frac{TN}{TN + FP} \tag{27}$$

The optimal operating point for the classifier in the ROC space would be obtained at the upper left point (0, 1). This point denotes the high accuracy where the sensitivity and specificity have their maximum values. Points falling under the diagonal line (which connects (0, 0) point to the (1, 1) point) denotes low sensitivity i.e. the classifier fails to detect the disease when it exists and low specificity i.e. the classifier misclassifies the disease when it does not exist. Area under the ROC curve (AUC) is another alternative for estimating the classifier accuracy. However, AUC mainly depicts the range of false positive

examples, while in medical diagnosis applications we desire the area with high sensitivity (true positive cases) [Paclik, 2008].

3. Results

This section represents the best classification results which are achieved under different conditions with SVM and *k*NN classifiers using various error estimation methods. In all classification tests performed in this work (considering a 2-class problem) the potentially precancerous lesions (OLP/OSF class) were labeled as {+1} and the healthy cases (NOE class) as {-1}. We implemented the experiments in Matlab environment and used Steve Gunn's toolbox in case of SVM classifier [Gunn, 1998].

3.1. Feature Selection

To select a subset of features for improved discriminatory ability, we performed a wrapper approach by using forward selection algorithm for all possible subsets consisting of one to eight features. The best classification accuracy 83.7% (135 corrected classified samples out of 161) was observed for the subset of four features (including entropy, mean, skewness, and variance). As shown in Figure 3.1, adding more features to the current subset even degraded in classification performance possibly due to the irrelevant or less discriminative features. The experiment was performed using the 5-fold cross validation error and *RBF*-SVM classifier (σ=0.5, C=100). The same experiment for sub-epithelial area returned roughly the same results (classification accuracy = 84%).

The scatter plot (a 2-dimensional representation of data distribution) for two of the most discriminative features (entropy and mean) is illustrated in Figure 3.2. The scatter plot implies that the precancerous samples build almost separated clusters whereas there is comparatively higher correlation between precancerous and healthy samples. This approves the physician's claim on considering the classification problem of healthy samples vs. the precancerous lesions as the most challenging diagnosis in the content of oral cavity diseases. To provide a better visualization perception of the data distribution and, moreover, to perceive how complex or simple the separation margin is, a 2-dimensional decision boundary is represented in Figure 3.3. The undesired island-like regions could be

the result of a few misclassified samples which disturb the clusters and lead to very complex boundaries.

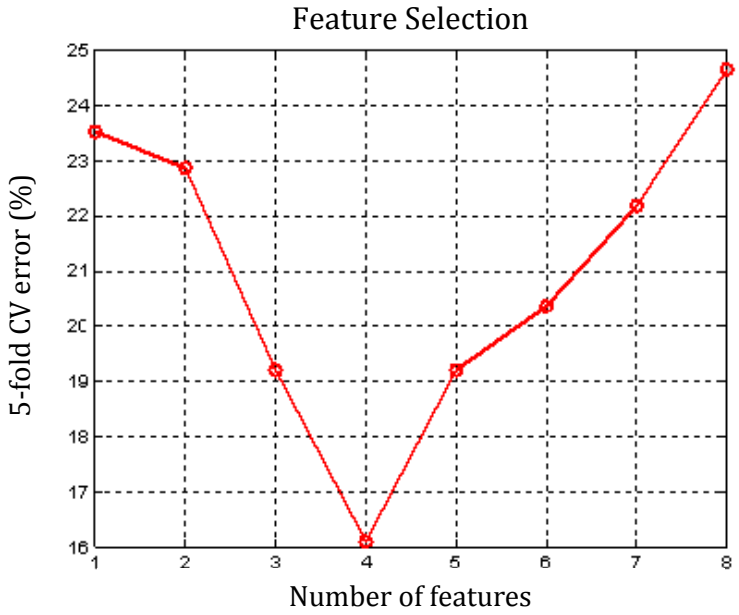


Figure 3.1: Best result obtained from 5-fold cross validation error rate vs. the number of histogram-based features, using RBF-SVM classifier (σ=0.5, C=100), two class problem (61 numbers of healthy samples vs. 100 numbers of premalignant lesions). The lowest error rate (16.3%) belongs to the subset of subset of four features (including entropy, mean, skewness, and variance)

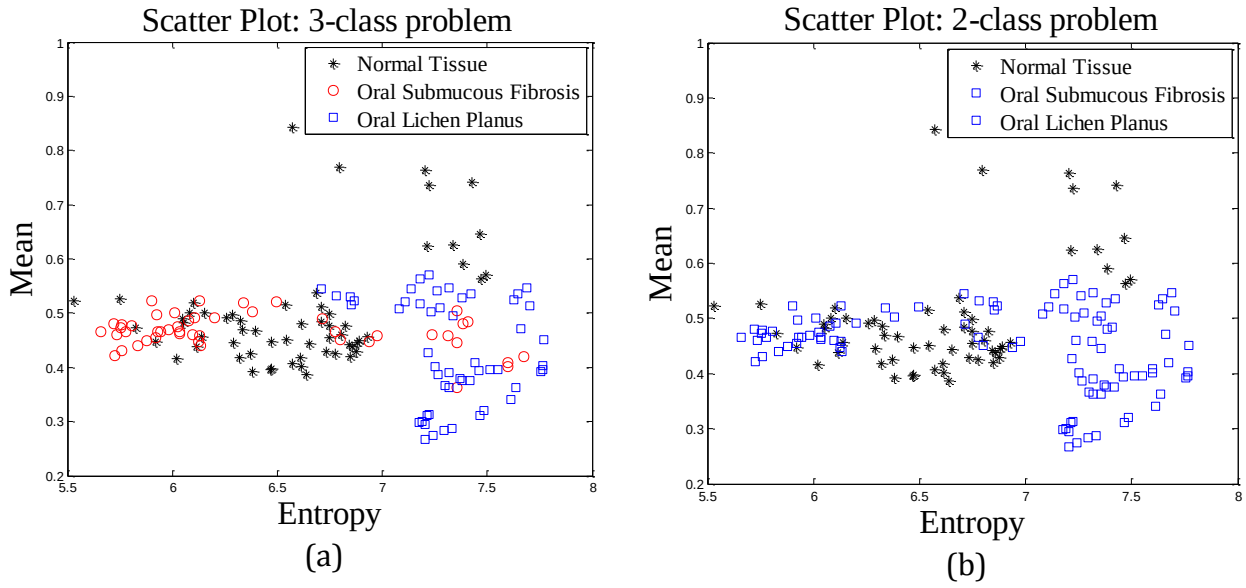


Figure 3.2: A typical scatter plot using two features (entropy and mean) for a) 3-class problem, b) 2-class problem (healthy samples vs. premalignant lesions). 161 samples are taken from juxta-epithelial area and classified using RBF-SVM (σ=0.5, C=100) classifier.

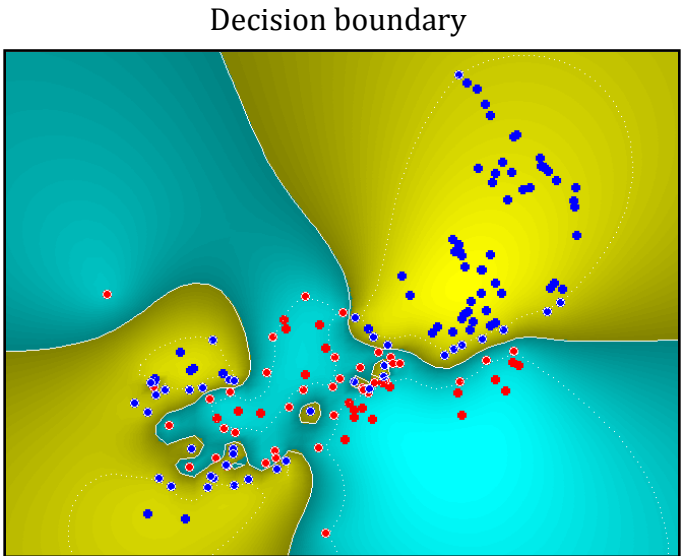


Figure 3.3: A typical decision boundary using 2 features (entropy and mean), exponential rfb-svm classifier (σ=1), 2-class problem (healthy samples vs. premalignant lesions), 161 samples, juxta-epithelial area, nsv (number of support vectors) = 79(49.1%).

3.2. Classifier Design

3.2.1. Estimation of "*C*" Value

One cannot know beforehand which parameter C (eq. 18) will be the best for a classification problem. So, we did a grid search on C, using all available samples (161) extracted from juxta epithelial area, to identify the one with the best prediction accuracy. Three alternative error estimation methods, 5-fold cross validation, resubstitution and leave-one out, have been used (Figure 3.4).

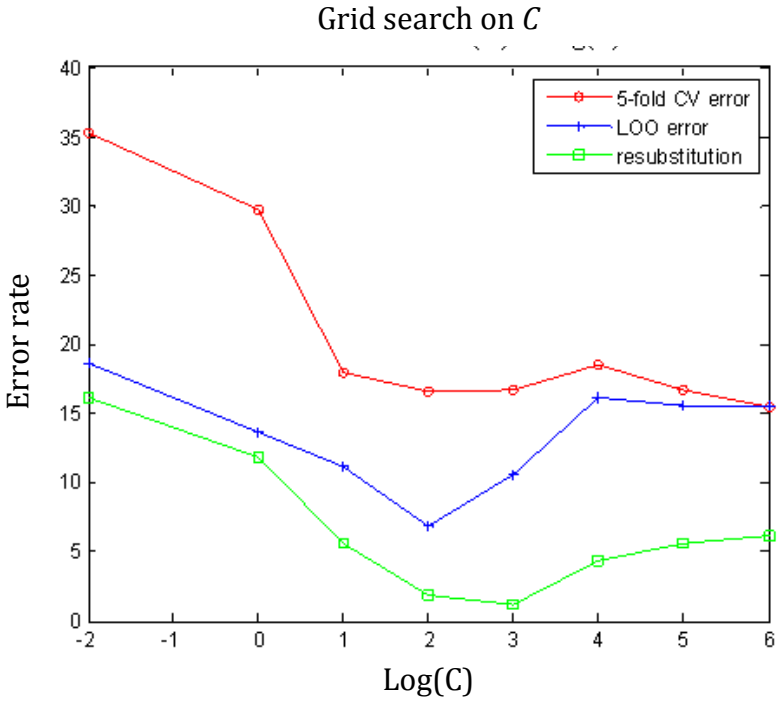


Figure 3.4: Error measures versus penalty parameter *C* based on the different error rate estimations. RBF-SVM: σ=1, dim=4 (features: entropy, mean, skewness, and variance), juxta-epithelial area, two class problem (61 NOM cases vs. 100 OSF/OLP lesions)

Figure 3.4 suggests that the resubstitution error can possibly behave optimistically for this classification problem (since all samples are devoted for the classifier design, so the error does not measure any unseen data). Thus, this measure could not be sufficient alone to judge on. Based on the LOO-error and 5-fold cross-validated error, the best classification

accuracy was achieved with C=10^2. However, 5-fold curve also had another minimum corresponding to C=10^6.

3.2.2. "Kernel" Selection

Once the penalty parameter C was chosen, we investigated for the right kernel function (eq. 16). To date, there is no theoretical method to find an appropriate kernel function for a particular problem. Therefore, we randomly evaluated a few kernels (mentioned in Table 2.2) with varying their kernel parameters e.g. kernel width (σ) and polynomial degree (d). The average SVM results for kernel parameter versus penalty term C are presented as percentage in Table 3.1.

Table 3.1: Choice of kernel function based on the various penalty parameters C, polynomial degree (d), and kernel width (σ) for SVM classifier, dim=4, two class problem (61 NOE samples vs. 100 OSF/OLP lesions), and 5-fold cross validation error

Log(C)	**-2**		**0**		**1**		**2**		**4**		**6**	
Kernel	nsv	Err	nsv	Err	nsv	Err	nsv	Err	nsv	Err	nsv	Err
Linear	99	**38**	91	**43**	86	**45**	85	**45**	85	**45**	85	**45**
RBF σ=0.5	129	**27**	87	**25**	66	**23**	58	**27**	56	**27**	33	**27**
RBF σ=1	129	**35**	78	**30**	54	**18**	40	**17**	55	**18**	39	**15**
RBF σ=2	129	**45**	91	**35**	62	**32**	46	**30**	106	**25**	48	**37**
Poly d=2	95	**30**	59	**33**	48	**33**	92	**32**	15	**35**	3	**35**
Poly d=3	87	**31**	48	**21**	49	**19**	75	**15**	12	**44**	3	**44**
Poly d=4	75	**33**	41	**17**	53	**22**	62	**23**	38	**22**	10	**32**
Poly d=5	65	**26**	48	**22**	49	**22**	48	**23**	29	**22**	9	**22**

From Table 3.1 we observed that the minimum classification errors were obtained with *RBF*-SVM and kernel width σ=1. The other kernels produced relatively higher error rates. In case of polynomial kernel, for a constant C value, higher polynomial degrees returned lower error rates and lower number of support vectors (nsv). However, for constant polynomial degrees, some minimum was achieved in the middle range of the parameter C (e.g. 10^2); meaning that an average value of C could be appropriate to define the separating hyper plane. Once the parameter C and kernel function are determined, the classifier can be designed based on the identified optimal hyper plane.

3.3. Classification Results

Intended for a practical comparison between different classifiers and also to evaluate the performance of classifiers, we constructed the Receiver Operating Characteristic (ROC) curves. By varying the bias *b* term in eq. 15 some operating points on ROC plane (False Positive rate (*FPr*, 1-specificity) versus True Positive rate (*TPr,* sensitivity)) were obtained. The final ROC curve is actually composed of the mean *TPr* and *FPr* trials.

3.3.1. ROC Curves Using Histogram-Based Features

Figure 3.5 represents ROC curves for three classifiers. 5-fold CV error, resubstitution error and all data samples from juxta-epithelial area have been used. We observed that the ROC curve for resubstitution error was over-optimistic and possibly resulting in an over trained classifier. The ROC curves for the two other classifiers showed relatively the same performance as they both represented roughly 90% sensitivity at specificity around 70%. However, rfb-SVM (*C*=100) performed slightly better than *k*NN. It showed more increase in sensitivity with decreasing the specificity and reached the operating point with 100% sensitivity earlier than *k*NN. Nevertheless, most likely because of the insufficient number of the samples in our database, this performance difference between the two classifiers cannot be considered as a statistically significant difference [Paclik, 2008]. The total classification accuracy reached a level of 81.4% (131 out of 161, 5-fold CV) compared with the regular way of measuring classification accuracy (counting the number of misclassified cases and averaging the result) that was 83.7%.

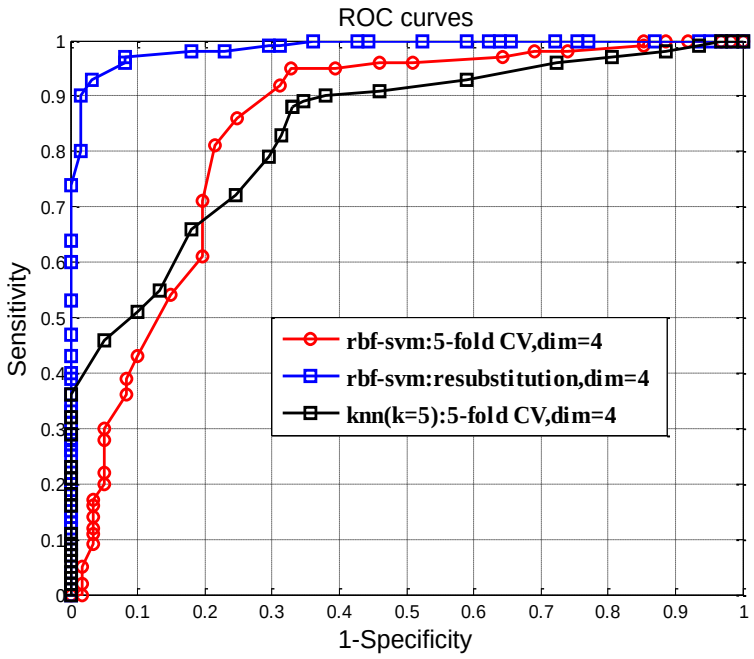
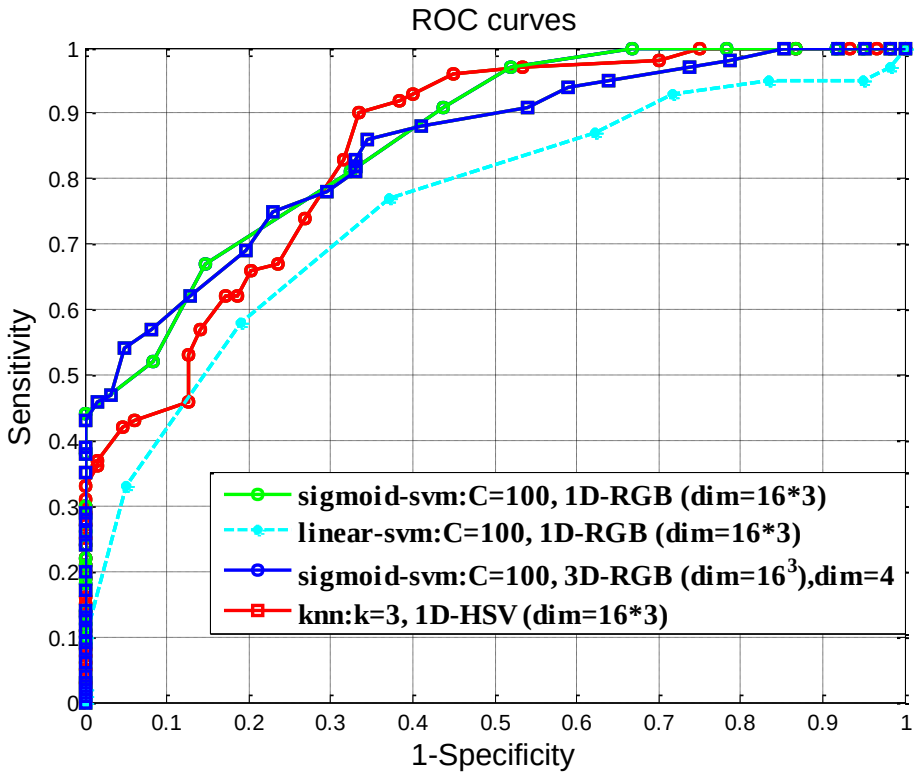


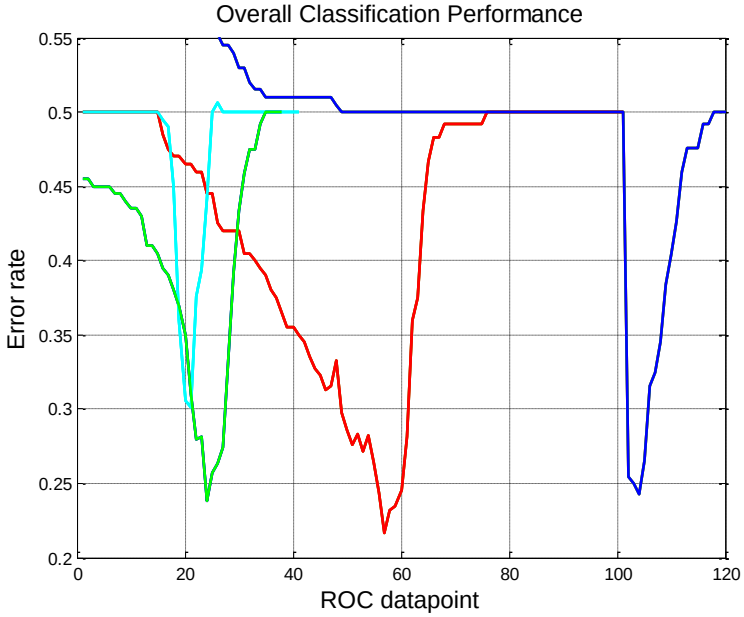
Figure 3.5: Receiver Operating Characteristic (ROC) curves for *k*NN classifier (*k*=5), RBF-SVM classifier (σ=1, C=10^2), dim=4 (selected features: entropy, mean, skewness, and variance), 5-fold cross validation, and resubstitution error. Two-class problem: NOM (N1=61) vs. OSF/OLP (N2=100) for juxta-epithelial area.

3.3.2. ROC Curves Using Color Histograms

For the sake of a new experiment, it was of interest to investigate the performance of the classifier built on the raw histogram bins served as input features to the classifier. We proposed to investigate the discrimination power of the two commonly used histograms, RGB and HSV color spaces. Considering 16 bins per color channel, we got a feature vector of 4096 entries (or 4096 partitions on the color space) for three-dimensional histograms (16^3), and 48 entries for one-dimensional histograms (16·3). Figure 3.6(a) illustrates the ROC curves for one- and three-dimensional RGB and HSV histograms using different classifiers. Their corresponding diagrams representing the total classification performance are shown in Figure 3.6(b).



(a)



(b)

Figure 3.6: a) ROC curves for: 1D-RGB histogram (dim=16*3) using sigmoid-SVM (ρ=1, C=100) and linear-SVM (C= 100) classifiers; 3D-RGB histogram (dim=163) using sigmoid-

SVM (ρ=4, C=100); 1D-HSV histogram (dim=16*3) using kNN classifier (k=3), 5-fold cross validation error for cancer vs. non-cancer classification problem. b) Overall classification performance corresponding to the ROC curves represented in (a).

Figure 3.6(a) showed that the proposed SVM classifiers, in comparison with *k*NN, return slightly better classification accuracy by using the original one- and three-dimensional RGB histograms. This might be explained by the high generalization performance of SVM classifier even with the large dimension of feature vector. On the other hand, HSV system provided almost the same discrimination ability for both SVM and *k*NN classifiers. However *k*NN is preferred as it is easier to implement and is less computationally expensive. The precise classification error rates versus the number of ROC data points for the examined classifiers are shown in Figure 3.6(b). The error rates were calculated as follows:

$$\varepsilon_{total} = \rho_1 \varepsilon_1 + \rho_2 \varepsilon_2 \tag{28}$$

where $\hat{\varepsilon}_i = \frac{\hat{\tau}_i}{N_i}$ as defined in eq. 22 is the error rate in which τ_i denotes the number of misclassified unknown samples (FPr+FNr) and N_i denotes the number of test samples. ε_1 and ε_2 are FP (false positive) and TP (true positive) rates, respectively. In order to produce different error rates, we weighted the error rate ε_i by multiplying in e.g. $\rho_1 = c$ and $\rho_2 = 1 - c$ where c is a constant value. The maximum classification rate of 78 percent (as suggested in Figure 3.6(b)) was achieved with *k*NN classifier in HSV system.

4. Conclusion & Future Work

4.1. Conclusion

- Generally, Better diagnostic performance obtained when using histogram-based features than RGB/HSV color systems as discriminative features to the classifiers.

- The overall diagnostic performance of 84% obtained with the proposed classification method in this work might be compared with the human visual classification rate of 74% [Jullien, 1995]. So the proposed system can be considered as a supportive tool to be used with general physicians/dentists in distinguishing the pre-cancerous oral conditions (OSF/OLP) against the healthy tissues.

4.2. Future work

- In this paper, we have investigated two commonly used classifiers; Support Vector Machines (SVM) and *k*-Nearest Neighbors (*k*NN). However, there are many other choices of classifiers worthy of testing. Especially as the type of classifier highly influences the classification performance, it would be of interest to try some other types such as Artificial Neural Networks (ANN) or Bayes Classifiers.

- The extracted features are of great importance and will have the most impact on the diagnostic performance regardless of which type of classifier has been used. Here, we have investigated the discrimination power of two color systems, RGB and HSV, and some histogram-based properties of tissue. However, it would be of interest to repeat the same experiment with a new set of possibly more potential features. It should be mentioned that the discrimination power of features is highly dependent on the distribution of data. So, it would be difficult to draw any

definite conclusion on which set of features would perform optimally in one particular classification problem.

- The segmentation of oral lesions has been manually performed in the current work. An automatic segmentation approach, however, can speed up the overall procedure.

- In order to provide a better comparison between classifiers, a larger volume of samples is required.

5. References

[Abbey, 1997]
Abbey L.M., Kaugars G.E., Gunsolley J. C., Bums J. C., Page D.G., Svirsky J. A., Eisenberg E., and Krutchkoff D. J., "The effect of clinical information on the histopathologic diagnosis of oral epithelial dysplasia", *Virgina Commonwealth University,* October 1997

[Abbey, 1995]
Abbey L.M., Kaugars G.E., Gunsolley J.C., Burns J.C., Page D.G., Svirsky J.A., Eisenberg E., and Krutchkoff D. J., "Intraexaminer and interexaminer reliability in the diagnosis of oral epithelial dysplasia", *Oral Surg Oral Med Oral Pathol Oral Radiol Endod,* August 1995

[AHNS, 1998]
AHNS Committee, "Oral Cavity Cancer", the Global Site: American Head and Neck Society, 1998 from
http://www.ahns.info/

[Axell, 1976]
Axell T., "A prevalence study of oral mucosal lesions in an adult Swedish population"*, odontologisk Revy*, Vol. 27, pp. 52-54, 1976

[Axelberg, 2007]
Axelberg P. G. V., Gu I. Y. H., and Bollen M. H. J., "Support Vector Machine for Classification of Voltage Disturbances"*, IEEE Transactions on Power Delivery*, Vol. 22, No. 3, July 2007

[Brickely, 1996]
Brickely M. R., Gospe J. H., and Shepherd J. P., "Performance of a computer simulated neural network trained to categorise normal, premalignant and malignant oral smears"*, Oral Pathol Med ,* pp. 424-428, 1996

[Chan, 2004]
Chan H. P., Sahiner B., and Hadjiiski L., "Sample Size and Validation Issues on the Development of CAD Systems"*, Computer Assisted Radiology and Surgery. Proceedings of the 18th International Congress and Exhibition,* Vol. 1268, pp. 872-877, June 2004

[Chapelle, 1999]
Chapelle O., Haffner P., and Vapnik V. N., "Support Vector Machines for Histogram-Based Image Classification", *IEEE Transactions on Neural Networks,* Vol. 10, No. 5, September 1999

[Chih-Wei, 2009]
Chih-Wei H., Chih-Chung C., and Chih-Jen L., "A Practical Guide to Support Vector Classification", *National Taiwan University, Taipei 106, Taiwan,* May 2009

[Chodorowski, 2009]
Chodorowski A., Hazarey V., and Kothawar S. K., "Pattern Toward Automated Classification of Oral Lesions from Histological Images", 2009

[Chodorowski, 2000]
Chodorowski A., "Pattern Recognition Methods for Oral Lesion Classification using Digital Color Images", *Departments of Signals and Systems*, Chalmers *University of Technology*, Goteborg, Sweden, 2000

[Cortes & Vapnik, 1995]
Cortes C., and Vapnik V. N., "Support Vector Networks"*, Machine Learning*, Vol. 20, pp. 237-297, 1995

[Christodoulou, 2003]
Christodoulou C. I., Michaelides S. C., and Pattichis C. S., "Multiple Texture Analysis for the Classification of Clouds in Satellite Imagery"*, IEEE Transactions on Geossience and Remote Sensing,* Vol. 41, No. 11, November 2003

[Epstein, 2002]
Epstein J. B., Zhang L., and Rosin M., "Advances in the Diagnosis of Oral Premalignant and Malignant Lesions", *Journal of the Canadian Dental Association*, Vol. 68, No. 10, November 2002

[Fukunaga & Hayes (a), 1989]
Fukunaga K., and Hayes R. R., "Estimation of Classifier Performance"*, IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 10, October 1989

[Fukunaga & Hayes (b), 1989]
Fukunaga K., and Hayes R. R., "Effects of Sample Size in Classifier Design", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 8, August 1989

[Fukunaga & Hayes, 1990]
Fukunaga K., and Hayes R. R., "Introduction to Statistical Pattern Recognition"*, Academic Press Professional*, 1990

[Gunn, 1998]
Gunn S. R., "MATLAB Support Vector Machine Toolbox"*, Image Speech and Intelligent Systems Research Group, University of Southampton,* March 1998

[Hsu, 2009]
Hsu C. W., Chang C. C., and Lin C. J., "A Practical Guide to Support Vector Classification", *Department of Computer Science, National Taiwan University,* May 2009

[Jafari-Khouzani, 2005]
Jafari-Khouzani K., and Soltanian-Zadeh H., "Radon Transform Orientation Estimation for Rotation Invariant Texture Analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 27, No. 6, June 2005

[Jain & Zongker, 1997]
Jain, A. K. and Zongker, D., "Feature Selection: Evaluation, Application and Small Sample Performance", *IEEE Trans. PAMI,* Vol. 19, pp. 153-158, 1997

[Jakkula]
Jakkula V., "Tutorial on Support Vector Machine (SVM)", *School of EECS, Washington State University*

[Jullien, 1995]
Jullien J. A., Downer M. C., Zakrzewska J. M., and Speight P. M., "Evaluation of a screening test for the early detection of oral cancer and precancer", *Community Dent Health,* Vol. 12, pp. 3-7, March 1995

[Kanwar, 2009]
Kanwar P., "Oral Cancer – Statistics, Risk factors, Signs and Symptoms, Diagnosis and Treatment", February 2009 from
http://www.xomba.com/

[Kim, 2006]
Kim H. D., Park C. H., Yang H. C., and Sim K. B., "Generic Algorithm Based Feature Selection Method Development for Pattern Recognition", *International Joint Conference SICE-ICASE*, pp. 1020-1025, October 2006

[Liu, 2009]
Liu W., and Wang Y., "A Target Detection Algorithm Based on Histogram Features and Particles", *Fifth International Conference on Natural Computation, China University of Petroleum,* pp. 206-209, 2009

[Liu, 2005]
Liu H., Dougherty E.R., Dy J.G., Torkkola K., Tuv E., Peng H., Ding C., Long F., Berens M., Parsons L., Zhao Z., Yu L., and Forman G., "Evolving Feature Selection", *IEEE Intelligent Systems*, Vol. 20, pp. 64-76, November-December 2005

[Lountzis (a), 2009]
Lountzis N. I., "Oral Submucous Fibrosis", *Department of Dermatology and Pathology, Geisinger Medical Center,* March 2009

[Lountzis (b), 2009]
Lountzis N. I., "Health Consequences of Tobacco Use: Tobacco-related Oral Mucosal Lesions and Dental Diseases", *Department of Dermatology and Pathology, Geisinger Medical Center,* March 2009

[MATLAB, 2004]
MATLAB Image Processing Toolbox User's Guide, *the Mathworks Inc., Natick, Massachusetts,* 2004 from
http://www.mathworks.com/

[Mazurowski, 2007]
Mazurowski M. A., Habas P A., Zurada J. M., Lo J. Y., Baker J. A., and Tourassi G. D., "Training Neural Network Classifiers for Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance", *Neural Networks,* Vol. 21, pp. 427-436, March 2007

[MFMER, 1998]
MFMER, "Oral Lichen Planus", *the Global Site: Mayo Foundation for Medical Education and Research*, 1998 from
http://www.mayoclinic.com/

[Miyamoto, 2003]
Miyamoto T., Uchimura S., Hamamoto Y., Jizuka N., Oka M., and Yamada-Okabe H., "Comparative Study of Feature Selection Methods on Microarray Data", *IEEE EMBS Asian-Pacific Conference on Biomedical Engineering,* pp. 82-83, October 2003

[NCI, 2009]
NCI Committee, "General Information about Lip and Oral Cavity Cancer", *the Global Site: National Cancer Institute, U.S. National Institutes of Health*, 2009 from
http://www.cancer.gov/

[Paul, 2005]
Paul R. R., Mukherjee A., Dutta P. K., Banerjee S., Pal M., Chatterjee J., Chaudhuri K., and Mukkerjee K., "A novel wavelet neural network based pathological stage detection technique for an oral precancerous condition", *Journal of Clinical Pathology*, Vol.58 pp.932-938, September 2005

[Pearson, 1895]
Pearson K., "Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material", *Philosophical Transactions of the Royal Society of London*, pp. 343-414, 1895

[Paclik, 2008]
Paclik P., Lai C., Novovicova J., and Duin R. P. W., "Variance estimation for two-class and multi-class ROC analysis using operating point averaging", *19th International Conference on Pattern Recognition,* pp. 1-4, 2008

[Qin, 2005]
Qin Z. C., "ROC Analysis for Predictions Made by Probabilistic Classifiers", *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou,* pp. 18-21, August 2005

[Rhodus, 2003]
Rhodus N. L., Myers S., and Kaimal Sh., "Diagnosis and Management of Oral Lichen Planus", *Northwest Density; Journal of the Minnesota Dental Association,* Vol. 82, No. 2, March-April 2003

[Sahiner, 2007]
Sahiner B., Chan H. P., and Hadjiiski L., "Classifier Performance Estimation under the Constraint of a Finite Sample Size: Resampling Schemes Applied to Neural Network Classifiers"*, Neural Networks,* Vol. 21, pp. 476-483, 2007

[Sergyan, 2008]
Sergyan S., "Color Histogram Features Based Image Classification in Content-Based Image Retrieval Systems", *6th International Symposium on Applied Machine Intelligence and Informatics, Institute of Software Technology, Budapest ,* pp. 221-224, 2008

[Sewell, 2007]
Sewell M., "Feature Selection"*, IEEE Intelligent Systems*, 2007

[Song, 2007]
Song Y., Huang J., Zhou D., Zha H., and Giles C. L., "IKNN: Informative K-Nearest Neighbor Pattern Classification", *Department of Computer Science and Engineering, College of Information Sciences and Technology*, *USA*, pp. 248-264, 2007

[Sugerman, 2002]
Sugerman P. B., "Oral Lichen Planus: Causes, diagnosis and management", *Australian Dental Journal*, AstraZeneca R&D Boston, Vol. 47, pp. 290-297, January 2002

[Surak, 2002]
Surak S., Qian G., and Pramanik S., "Segmentation and Histogram Generation Using HSV Color Space for Image Retrieval", *International Conference on Image Processing,* Vol. 2, pp. 589-592, 2002

[Swain, 1990]
Swain M.J., and Ballard D.H., "Indexing Via Color Histograms", *Third International Conference on Computer Vision,* pp. 390-393, 1990

[Vandewalle, 2003]
Vandewalle J., and Van Huffel S., "Least Squares Support Vector Machines Classification Applied to Brain Tumor Recognition Using Magnetic Resonance Spectroscopy", December 2003

[Vapnik, 1999]
Vapnik V. N., "An Overview of Statistical Learning Theory", *IEEE Transactions on Neural Networks*, Vol. 10, No. 5, September 1999

[Varma, 2006]
Varma S., Simon R., "Bias in error estimation when using cross-validation for model selection", *BMC Bioinformatics*, *Biometric Research Branch, National Cancer Institute, USA*, February 2006

[Yuan, 1999]
Yuan H., Tseng S. S., Gangshan W., and Fuyan Z., "A Two-phase Feature Selection Method using both Filter and Wrapper", *IEEE SMC '99 Conference Proceedings, IEEE International Systems, Man and Cybernetics*, Vol. 2, pp. 132-136, 1999

[Zerdoner, 2003]
Zerdoner D., "The Ljubljana classification- its application to grading oral epithelia hyperplasia", *Journal of Cranio-Maxillofacial Surgery, Hospital Celje, Slovenia,* pp. 75-79, April 2003