

IDENTIFICATION AND CLASSIFICATION OF ORAL CANCER LESIONS IN COLOR IMAGES

A DISSERTATION

*Submitted in partial fulfillment of the
requirements for the award of the degree*

of

MASTER OF TECHNOLOGY

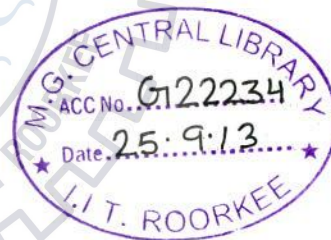
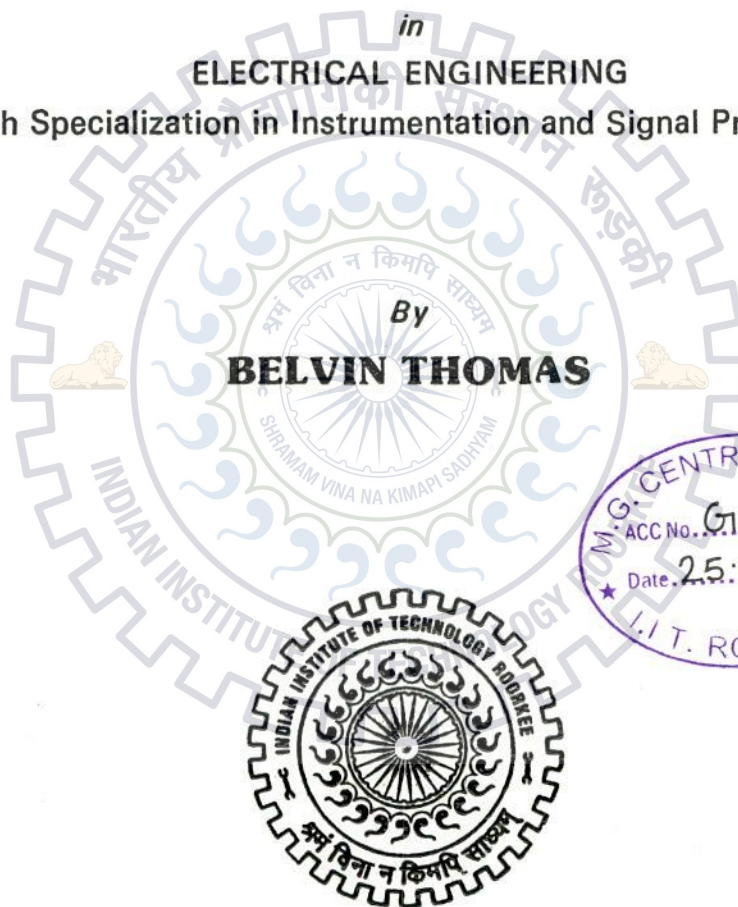
in

ELECTRICAL ENGINEERING

(With Specialization in Instrumentation and Signal Processing)

By

BELVIN THOMAS



**DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE - 247 667 (INDIA)**

JUNE, 2013

CANDIDATE'S DECLARATION

I hereby declare that this thesis report entitled "Identification and Classification of Oral Cancer Lesions in Color Images", submitted to the Department of Electrical Engineering, Indian Institute of Technology, Roorkee, India, in partial fulfillment of the requirements for the award of the Degree of Master of Technology in Electrical Engineering with specialization in Instrumentation and signal Processing is an authentic record of the work carried out by me during the period from September 2012 to May 2013 under the supervision of **Prof. VINOD KUMAR, Department of Electrical Engineering, Indian Institute of Technology, Roorkee.** The matter presented in this thesis report has not been submitted by me for the award of any other degree of this institute or any other institute.

Date: 27/5/13

Place: Roorkee


Belvin Thomas

CERTIFICATE

This is to certify that the above statement made by the candidate is true to the best of my knowledge and belief.


Prof. VINOD KUMAR

Professor & Dean of Faculty Affairs
Department of Electrical Engineering
IIT Roorkee

"Science may set limits to knowledge, but should not set limits to imagination."

Bertrand Russell



ABSTRACT

Oral Cancer is the most common form of cancer in India. Poor adult villagers from remote areas in their 60s and 70s are the usual victims of oral cancer. The peculiar nature of oral cancer is that it is curable if treated properly at right time. In India this is particularly deadly due to the extensive use of tobacco coupled with lack of proper diagnostic facilities. A biopsy done at right time could save many lives. But it is the ignorance regarding the right time that causes the consequent sufferings and unavoidable death sentence. This happens since the symptoms appear to be very much similar to a normal mouth ulcer and resulting negligence and complacency. This underscores the need for easier methods to identify this villain in disguise at an early stage.

Early diagnosis via mass screening initiatives in rural areas with the help of image classification systems can help to reduce the mortality, medical costs, pain and sufferings. True color images are the most easily available input to such a system. The separation of mouth cavity images into cancerous and non-cancerous is a three stage process involving segmentation (Identification of ROI), feature extraction and classification. The objectives of this work include:

1. Study of different methods for segmentation, feature extraction and classification.
2. Selection of methods suitable for the present purpose.
3. Application of the selected methods on the database. The test database is formed by sample images acquired from the Medical College under HIHT, Dehradun.
4. Proposing the best system suitable for classification of images into cancerous and non-cancerous.

Acknowledgement

I would like to express my deep sense of gratitude and sincere thanks to my beloved guide **Prof. VINOD KUMAR**, Department of Electrical Engineering, Indian Institute of Technology Roorkee, for being helpful and a great source of inspiration. His keen interest and constant encouragement gave me the confidence to complete my work. I wish to extend my sincere thanks for his excellent guidance and suggestions for the successful completion of my work.

My sincere thanks to the Head, Electrical Engineering department, I.I.T. Roorkee for providing necessary research facilities to carry out this work and valuable suggestions and motivations provided by other faculty members of the department are duly acknowledged.

Special thanks to **Dr. Sunil Saini**, Director, CRI, HIHT, Dehradun for the set of images and the support provided in the form of domain knowledge transfer. His societal commitment serves as a major motivation for me.

I would like to extend my gratitude to my friends and seniors for their support throughout the past two years. I deeply acknowledge the support and advice provided by the research scholar Mr. Jitendra Virmani during initial stages of my work.

I dedicate this thesis to my family who unremittingly supported me during my years of study. Their relentless backing made this work possible.

Belvin Thomas
(11528007)

...

Contents

| | |
|--|-----------|
| Candidate's Declaration | i |
| Abstract | iii |
| Acknowledgement | iv |
| List of Figures | viii |
| List of Tables | ix |
| Abbreviations | x |
| 1 Introduction | 1 |
| 1.1 Introduction | 1 |
| 1.2 The Problem Statement | 2 |
| 1.3 Oral Cancer- Facts and figures | 3 |
| 1.4 Thesis Route Map and Overview | 3 |
| 2 Literature Survey | 5 |
| 3 Database and Methodology | 8 |
| 3.1 The Image Database | 8 |
| 3.2 Patch Extraction | 9 |
| 3.3 Proposed Methodology | 9 |
| 4 Segmentation and Feature Extraction | 11 |
| 4.1 Introduction | 11 |
| 4.2 Active Contour Without Edges | 12 |
| 4.3 Patch Nomenclature | 13 |
| 4.4 Patch Selection | 15 |
| 4.5 Feature Extraction | 15 |
| 4.5.1 GLCM Features | 17 |
| 4.5.2 GLRL Features | 17 |
| 4.5.3 Intensity Based First order Statistical Features | 18 |
| 5 Feature Selection and Analysis | 19 |
| 5.1 Introduction | 19 |

| | | |
|----------|--|-----------|
| 5.2 | Boxplots | 20 |
| 5.3 | Feature Selection – Method 1 | 21 |
| 5.4 | Feature Selection – Method 2 | 23 |
| 5.5 | Conclusion | 27 |
| 6 | Classification | 28 |
| 6.1 | Introduction | 28 |
| 6.2 | Implementation | 28 |
| 6.2.1 | Classifier 1 BPANN | 28 |
| 6.2.2 | Classifier 2 SVM | 30 |
| 6.3 | Results and Discussion | 30 |
| 6.4 | Conclusion | 32 |
| 7 | Performance Analysis | 33 |
| 7.1 | Introduction | 33 |
| 7.2 | The Standard Methods | 34 |
| 7.2.1 | Principal Component Analysis (PCA) | 34 |
| 7.2.2 | Sequential feature selection | 35 |
| 7.2.3 | t-Test | 36 |
| 7.2.4 | Relief-F | 36 |
| 7.3 | Performance Parameters | 37 |
| 7.3.1 | Sensitivity | 37 |
| 7.3.2 | Specificity | 38 |
| 7.3.3 | Accuracy | 38 |
| 7.4 | Results and Discussion | 38 |
| 7.4.1 | PCA followed by LDA | 38 |
| 7.4.2 | SFS followed by SVM | 39 |
| 7.4.3 | t-Test followed by SVM | 40 |
| 7.4.4 | ReliefF followed by SVM | 41 |
| 7.5 | Conclusion | 42 |
| 8 | Multi-class Classification of Malignant Lesions Using ANN | 44 |
| 8.1 | Introduction | 44 |
| 8.2 | Image Database and methodology | 45 |
| 8.3 | Segmentation of ROI | 46 |
| 8.4 | Feature Extraction | 47 |
| 8.4.1 | GLCM Features | 47 |
| 8.4.2 | GLRL Features | 48 |
| 8.4.3 | Intensity Based First order Statistical Features | 48 |
| 8.5 | Feature Analysis and selection | 49 |
| 8.6 | Experimental Results | 55 |
| 8.7 | Conclusion | 57 |
| 9 | Conclusion and Future Scope | 59 |
| 9.1 | Conclusion | 59 |
| 9.2 | Future Scope | 61 |
| 9.3 | Awards and Recognition | 63 |

| | |
|-----------------|-----|
| <i>Contents</i> | vii |
| Publications | 64 |
| References | 65 |
| Image Database | 72 |
| Formulae Used | 74 |



List of Figures

| | | |
|-----|--|----|
| 1.1 | Potential areas of occurrence of oral cancer | 2 |
| 3.1 | Proposed Methodology | 10 |
| 4.1 | Output GUI of ROI Identification :Image No.2 | 13 |
| 4.2 | Output GUI of ROI Identification :Image No.14 | 14 |
| 4.3 | Segmented ROI overlaid on the original image | 14 |
| 5.1 | Sample Boxplot | 20 |
| 5.2 | Feature Selection – Method 1 | 21 |
| 5.3 | Boxplot of SRLGE(rejected by method 1) | 22 |
| 5.4 | Boxplot of SRE(selected by method 1) | 22 |
| 5.5 | Feature Selection – Method 2 | 24 |
| 5.6 | Boxplot of (d=3) diff. entropy (Rejected after step 5) | 25 |
| 5.7 | Boxplot of LGRE (selected after step5) | 25 |
| 5.8 | Boxplot of LGRE (Rejected after step 9) | 26 |
| 5.9 | Boxplot of RP (selected after step 9) | 26 |
| 7.1 | Classification Matrix : PCA + LDA | 39 |
| 7.2 | Classification Matrix : SFS + SVM | 40 |
| 7.3 | Classification Matrix : t-Test + SVM | 40 |
| 7.4 | Classification Matrix : ReliefF + SVM | 41 |
| 8.1 | Proposed Methodology for Six-class Classification | 46 |
| 8.2 | Boxplot of GLCM Energy (d=3) | 49 |
| 8.3 | Boxplot of GLCM SA (d=3) | 55 |
| 9.1 | Welcome screen of GUI | 61 |
| 9.2 | GUI in action | 62 |
| 9.3 | Screenshot of Letter of Appreciation | 63 |

List of Tables

| | | |
|-----|--|----|
| 4.1 | Features Extracted | 16 |
| 5.1 | Features Selected by Method 1 | 23 |
| 5.2 | Features Selected by Method 2 | 26 |
| 6.1 | Classification Matrices | 31 |
| 6.2 | Classification accuracy based on method of feature selection | 32 |
| 7.1 | Features Selected by SFS | 39 |
| 7.2 | Features Selected by t-Test | 41 |
| 7.3 | Features Selected by ReliefF | 42 |
| 7.4 | Performance Analysis Table | 43 |
| 8.1 | Six Groups of Malignant Cases | 45 |
| 8.2 | Feature Analysis Table | 51 |
| 8.3 | Six-class Classification Results | 57 |
| 8.4 | Classification Matrix for case 4 | 58 |
| 1 | First Order Statistical Features | 74 |
| 2 | GLCM Features | 75 |
| 3 | GLRL Features | 77 |

Abbreviations

| | |
|-------|--|
| ANN | Artificial Neural Network |
| BPANN | Back Propagation based Artificial Neural Network |
| CADx | Computer Aided Diagnosis |
| CRI | Cancer Research Institute |
| DE | Difference Entropy |
| FOS | First Order Statistics |
| GLCM | Grey Level Co-occurrence Matrix |
| GLN | Gray-Level Non-Uniformity |
| GLRL | Grey Level Run Length |
| GUI | Graphical User Interface |
| HGRE | High Gray-level Run Emphasis |
| HIHT | Himalayan Institute Hospital Trust |
| HSI | Hue Saturation Intensity |
| IDM | Inverse Ddifference Moment |
| KNN | K- Nearest Neighbour |
| LDA | Linear Discriminant Analysis |
| LGRE | Low Gray-level Run Emphasis |
| LRE | Long Run Emphasis |
| LRHGE | Long Run High Gray-level Emphasis |
| LRLGE | Long Run Low Gray-level Emphasis |
| MP | Mega Pixel |
| OCT | Optical Coherence Tomography |
| PCA | Principal Component Analysis |
| RLN | Run Length Non-Uniformity |
| ROI | Region Of Interest |

| | |
|---------------------|---|
| RP | Run Percentage |
| SA | Sum Average |
| SE | Sum Entropy |
| SFS | Sequential Forward Selection |
| SoS Variance | Sum of Squares Variance |
| SRE | Short Run Emphasis |
| SRHGE | Short Run High Gray-level Emphasis |
| SRLGE | Short Run Low Gray-level Emphasis |
| SV | Sum Variance |
| SVM | Support Vector Machines |



Chapter 1

Introduction

“If there is technological advance without social advance, there is, almost automatically, an increase in human misery.”

-Michael Harrington

1.1 Introduction

Image characterization using texture analysis is quite relevant in such cases where the surface under consideration forms sufficient number of distinguishable patterns which may or may not be recognized by a human observer. Needless to say, statistical approaches and computational methods are far better than human capabilities in deriving a conclusion based on analysis of such patterns [1]. In this work, the use of popular texture analysis methods based on digital image processing is proposed for building a computer aided diagnostic tool for oral cancer. The target is identification and classification of oral lesions into cancerous and normal based on the differences in superficial patterns and surface deformities caused by different cases of malignancy.

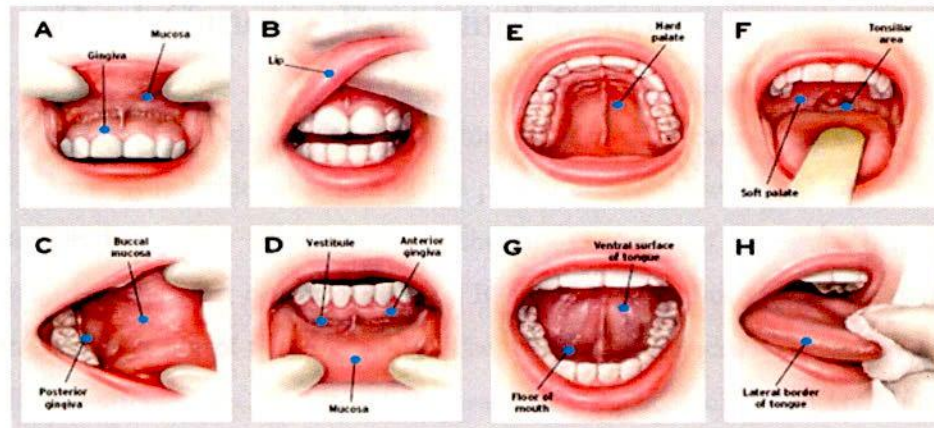


FIGURE 1.1: Potential areas of occurrence of oral cancer. [3]

1.2 The Problem Statement

Among all the methods of oral cancer diagnosis, a timely biopsy remains the golden standard[2]. But in a populous country like India it is not possible for each and every suspected victim to go for a biopsy. The reasons may be inaccessibility to a proper healthcare centre, lack of seriousness, lack of awareness or the large number of potential victims. Thus it is quite relevant in this scenario to develop an easier method which can be used in mass screening programmes. The capabilities of image analysis and soft computing methods can be used to make a simple system which can save many lives. True color images are the best possible input to such a system due to the ease with which they can be procured and processed these days. The precarious cases can be identified and classified from real-time images. In short, so far as oral cancer diagnosis is concerned, the proposed method of classification does not aim at replacing the importance of biopsy. But it will act as a supplement to the existing clinical methods in two respects.

1. To provide an early detection of malignancy and a timely advice to take up a biopsy.
2. To permit quantitative assessment during follow-up to the treatment by comparative studies.

1.3 Oral Cancer- Facts and figures

Cancer of the oral cavity is one of the highest occurring cancers in south and central Asia. Worldwide, only half of the people diagnosed with oral cancer live for another five years. Oral Cancer is the largest group of cancers which fall into the head and neck cancer category . It includes cancerous growth occurring within the structures of the oral cavity. The oral cavity includes teeth, lower jaw, vermilion borders of the lips, labial mucosae, buccal mucosae, buccal vestibulae, alveolar ridges and gingiva, floor of mouth, oral tongue, soft and hard palate 1.1. Common names are mouth cancer, tongue cancer, tonsil cancer or throat cancer. When found at early stages of development, oral cancers have an 80 to 90 percent survival rate [4]. India has high rate of incidence due to use of tobacco coupled with late diagnosis of potentially precancerous lesions.

1.4 Thesis Route Map and Overview

The major motivation behind this thesis is to propose a context specific method suitable for the development of an easier tool for oral cancer diagnosis at an earlier stage based on machine learning and computational approaches. Extensive study was conducted to understand various existing algorithms for classification and the current scenario in this field was analysed at the initial stages. The images were inducted into the database after discussions with medical domain experts and analysis was conducted with their help to develop understanding of the domain. ROI identification is performed using the best method suitable for image segmentation in the presented problem - curve evolution based Active Contour model.

Much importance is given for the idea of feature selection since it lies at the heart of any classification problem. A heuristic approach which is most suited for the given problem is proposed and implemented. The central idea behind the proposed approach is the hypothesis that malignancy can be easily identified with such features of the texture which manifest themselves with higher degree of similarity in case of normal images. An algorithm is developed to implement this idea and

the proposed methodology is validated with the help of two most popular classifiers - SVM and ANN. A performance analysis of the novel approach is conducted based on parameters like specificity, sensitivity and accuracy of the classification. Comparison is done with four of the existing methods for feature selection. The problem of classification of normal and malignant images is extended further towards a within class classification of malignant cases. Six ground-truth classes are identified and a multiclass classification is performed using ANN.



Chapter 2

Literature Survey

"Study the past if you would define the future"

-Confucius

From the existing literature regarding the methods for oral cancer identification and classification, it is clear that there is a need for better and easier approaches. Earlier attempts have seen the use of histopathological and microscopic images[5]. Color images were also used with color based features. Some of these techniques developed for early detection of oral cancer are based on advanced hardware while others are purely based on software.

A Chodorowski et al[6] suggested a method using true color images for oral lesion classification. Representations in five different color spaces were studied and their use in color image analysis of mucosal images was evaluated. Four common classifiers which were chosen for the evaluation of classification performance were (1)Fishers Linear Discriminant (2)Gaussian quadratic (3) KNN and (4) Multilayer perceptron. Resubstitution and Five fold cross validation methods were used for estimation of classification accuracy. By use of HSI color space and linear discriminant Analysis classifier, the best classification accuracy was achieved . A new color based method for automatic segmentation and classification of tumor tissues using microscopic images was proposed by Yung nien Sun et al [7] . The performance of the proposed fully automated method is compared against semi

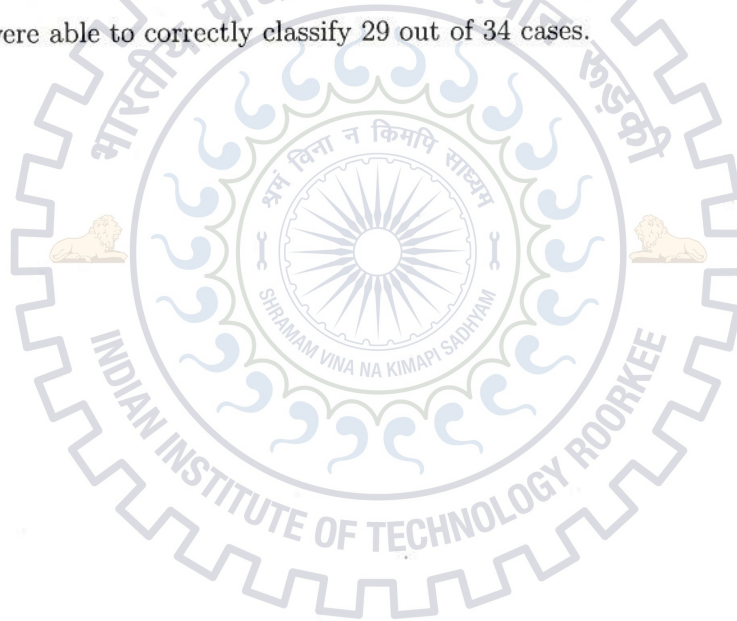
automated procedures. The proposed algorithm is developed into an effective tool to analyse oral cancer images. It is useful in analysis of other microscopic images with the same type of tissue staining.

Woonggyu Jung et al [8] proposed a technique for detection of oral cancer using Optical Coherence Tomography . OCT is suitable for oral mucosa for an imaging depth of 2-3 mm,. They were also able to detect oral cancer in 3-D images of oral lesions. Simon Kent [9] conducted a study and published a technical report on diagnosis of oral cancer using Genetic Programming. He provided a comparison between a Genetic Programming system and Neural Network model. Ranjan Rashmi Paul et al [10] proposed a methodology to detect oral cancer using wavelet - neural networks. The wavelet coefficients from microscopic images of collagen fibers of normal and malignant tissues have been used in order to constitute the feature vector which, in turn was used to train the Artificial Neural Network.

Ghassan Hamarneh, Tomas Gustavsson and Artur Chodorowski [11] proposed the use of a region based method for the segmentation of oral lesions in color images. They went on to perform a classification based on color based features. In this paper the researchers have also suggested the introduction of textural features in camera images to obtain better results. Yung nien Sun et al [12] proposed an automatic color based feature extraction system for parameter estimation of oral cancer from optical microscopic images. Comparison of Parameters between four stages of cancer was conducted. The proposed system is developed into a useful tool for automatic segmentation of stained biopsy samples of oral cancer.

The performance of data mining techniques for oral cancer prediction is compared by Neha Sharma, Nigdi Pradhikaran, Akurdi [13] . The two data mining techniques used are Multilayer Perceptron Neural Network model and tree Boost model. Tree Boost and Multilayer Perceptron Neural Network indicates the same specificity and sensitivity for Training data and validation data. There is no misclassification of data seen in both training and validation data in tree boost model and Multilayer Perceptron Neural Network. Also the most significant parameter for the cause of malignancy was identified to be the presence of Lymph Node. As per the study both the Multilayer Perceptron Neural Network model and Tree Boost Classification Model are suitable for predicting malignancy.

Chandran Chakraborty, M. Muthu Rama Krishnan, Ajoy Kumar Ray proposed a texture classification for oral histopathological sections based on the wavelet analysis [14]. A new method is proposed here since the conventional method involves higher misclassification error due to inter and intra observer variations. In the proposed method, feature extraction was done using wavelet transform. Kullback Leibler (KL) divergence was used for feature selection. Bayesian method and SVM were used for diagnostic classification. A computer aided diagnosis (CADx) system was designed by Arthur Chodorowski et al.[15] for oral mucosal lesions. They used medical cases from India as training samples. The classifiers used were SVM and Bayes point machine (BPM). The task was to discriminate between precancerous lesions and non-precancerous lesions. The discriminating features consisted of color differences and lesion shape properties. The observed classification accuracy was 85% for both SVM and BPM classification systems. They were able to correctly classify 29 out of 34 cases.



Chapter 3

Database and Methodology

"I have no data yet. It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts"

Sherlock Holmes

3.1 The Image Database

The material analyzed in this study was recorded at the Himalayan Institute of Medical Sciences under HIHT, Doiwala, Dehradun, India. The digital true color images were recorded using Sony Cybershot digital still picture camera (4x optical zoom, 10.1MP). The clinical diagnoses in malignant cases have been histopathologically verified. All photographs are acquired with 1X zoom position and the vertical and horizontal resolution is 72 dpi each. The contact person at the hospital is Dr. Sunil Saini, Director, Cancer Research Institute, HIHT, Dehradun. The entire database comprises of :

- 20 malignant images (20 persons)
- 51 normal images (10 persons)

3.2 Patch Extraction

Patches of size 32x32 are extracted from all the images in the database. The different sets of patches for the use in classification are finalized as given below:

- Malignant : 240 patches
- Normal : 240 patches

3.3 Proposed Methodology

Images acquired by a standard digital camera are given to a segmentation program which identifies the ROI. Patches of 32x32 size are extracted from each of the malignant ROI. Similar sized patches are extracted from normal images also. 61 features based on texture (GLCM), run-length (GLRL) and intensity variation are extracted. Suitable features are selected in two different methods with the help of boxplots. Finally two different classifiers (SVM and Back Propagation based Artificial Neural Network) are implemented. The behaviour of both the classifiers towards the different methods of feature selection is analysed. These classifiers will serve as the building stones for the potential oral cancer diagnostic tool. The Outline of the proposed methodology for classification of images into cancerous and non- cancerous is given in Fig. 3.1

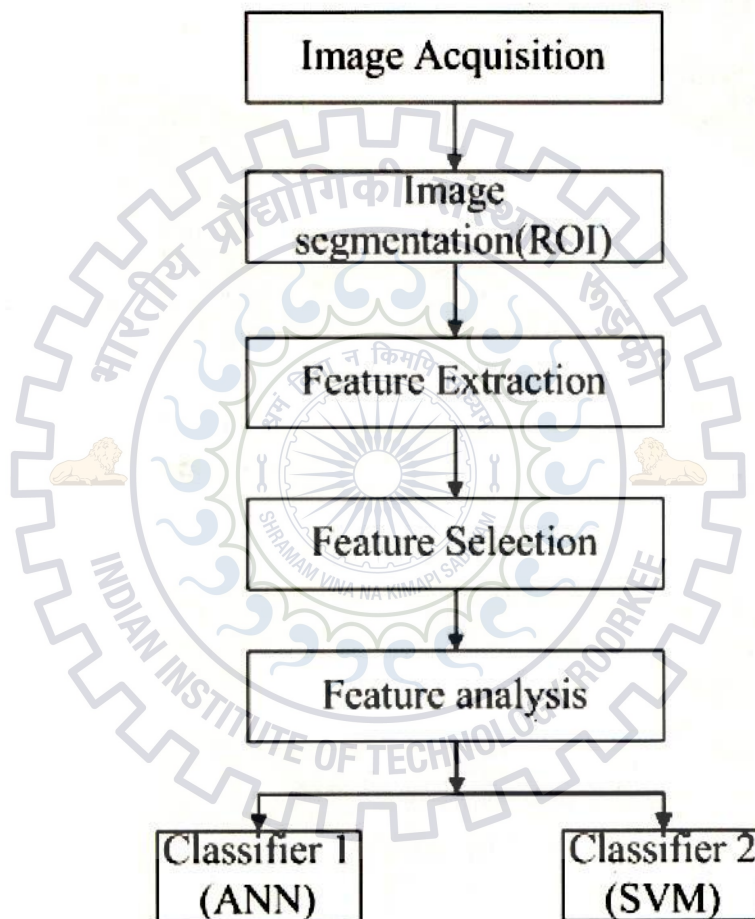


FIGURE 3.1: Proposed Methodology [16].

Chapter 4

Segmentation and Feature Extraction

"I am among those who think that science has great beauty. A scientist in his laboratory is not only a technician: he is also a child placed before natural phenomena which impress him like a fairy tale"



-Marie Curie

4.1 Introduction

Segmentation is the process of dividing an image into regions with similar properties such as grey level, colour, texture, brightness, and contrast. Thus image becomes more meaningful and easier to analyse. The role of segmentation is to subdivide the objects in an image. Segmentation occupies a very significant role in image processing and computer vision as it is the vital first step before subsequent steps like feature extraction, classification, image registration, reconstruction etc. Segmentation uses various algorithmic methods to convert standard imagery into information which is more helpful in analysis.

4.2 Active Contour Without Edges

This chapter is intended to explain the approach used for segmentation of the digital imagery used in the current problem. Active contour model without edges proposed by Tony F. Chan and Luminita A. Vese [17] is used here. It is a variation of classical active contour model to detect objects in a given image [18], based on curve evolution techniques, MumfordShah functional [19] for segmentation and level sets.[20] This model can detect objects whose boundaries are not necessarily defined by gradient. Energy minimisation is done which can be seen as a particular case of the minimal partition problem. In the level set formulation, the problem becomes a mean-curvature flow-like evolving the active contour, which will stop on the desired boundary. [21, 22] However, the stopping term does not depend on the gradient of the image, as in the classical active contour models, but is instead related to a particular segmentation of the image. Also, the initial curve can be anywhere in the image, and interior contours are automatically detected. The region of interest in the current problem consists of lesion area. The purpose here is to extract the lesion area from the rest of the image containing normal tissues. It was decided to use this particular method due to following reasons:

- It is a semi-automatic technique.
- It reduces the delineation time.
- It works well with all the test samples despite the high degree of variability involved.
- It is based on curve evolution and keeps on finding the boundary even though it is not well-defined by gradient.

The algorithm implemented is briefly enumerated below:

Step 1: An initial mask is created by selecting upper left and lower right points.

Step 2: A signed distance map (ϕ) is created from the mask.

Step 3: The curves narrow band is selected.

Step 4: The two terms of Energy are computed. [23]

(i) Based on Image Intensity Information

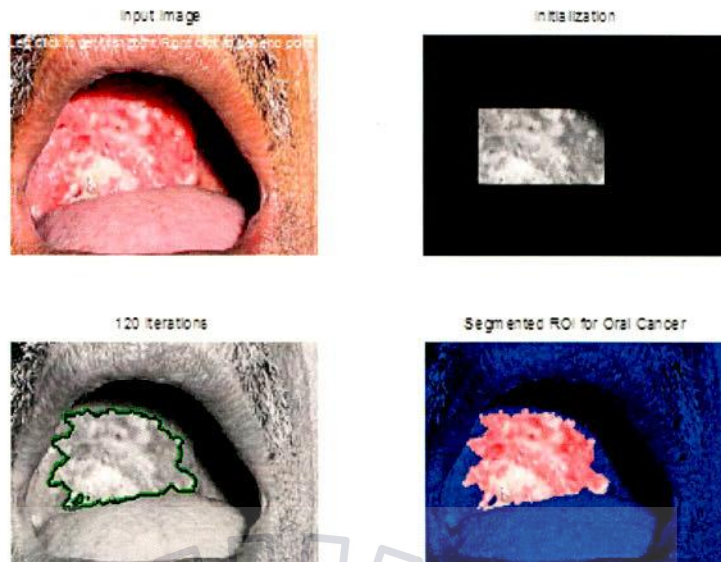


FIGURE 4.1: Output GUI of ROI Identification from image No.2 in database : Carcinoma Palate Ulceroproliferative growth. (i)Input image (upper left), (ii) Initial mask(upper right), (iii)Grey level image with boundary after 120 iterations(lower left), (iv)Segmented ROI for oral cancer (lower right).

(ii) Based on Curvature (Smoothness)

Step 5: Energy minimization by Gradient descent method.

Step 6: The new phi is computed and curve evolution is performed

Step 7 : Re-initialization of phi to keep Signed Distance Map smooth. [24, 25]

Step 8 : Display the intermediate output and go back to step 3. Continue this till the end of maximum number of iterations specified by the user.

Step 9: Segmented ROI is overlaid on the original image. Patches of 32x32 size are selected and saved in .tif format.

4.3 Patch Nomenclature

Naming convention followed for each patch extracted : `gxxxass_n`

Where

g- group(n=normal , m=malignant)

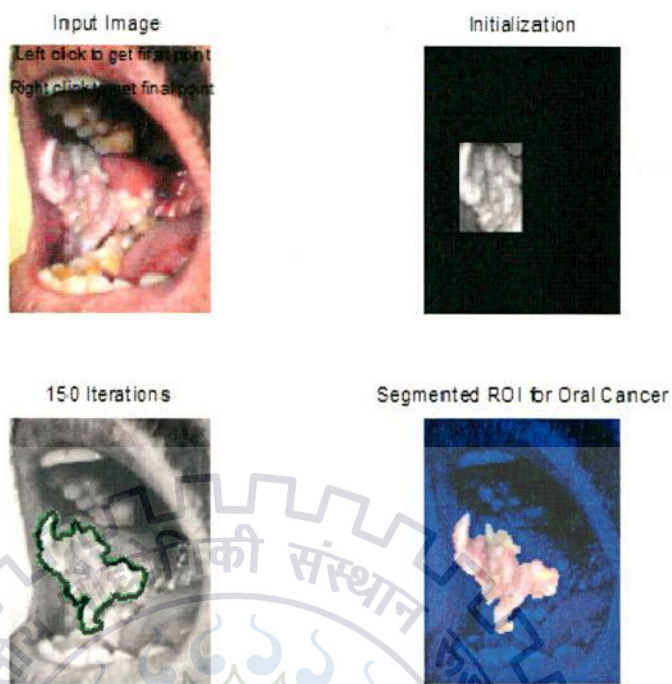


FIGURE 4.2: Output GUI of ROI Identification from image No.14 in Database : Verrucous carcinoma buccal mucosa. (i)Input image (upper left), (ii) Initial mask(upper right), (iii)Grey level image with boundary after 150 iterations(lower left), (iv)Segmented ROI for oral cancer (lower right).

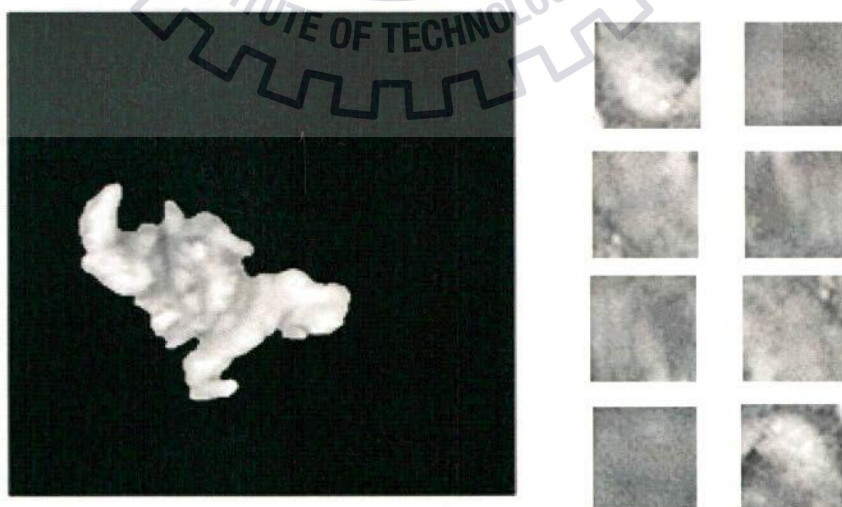


FIGURE 4.3: Segmented ROI overlaid on the original image with ROI in grayscale highlighted (left).Sample patches (right).

xxx- image no.(001,002,003.....)

a - Area of occurrence[u=upper ,l=lower ,b=buccal ,t=tongue ,o=outer(including lip)]

ss - Subject No. (only for normal images)

n - Patch No.(1,2,3,...)

Eg: m016t_10.tif

n003b01.2.tif

Such a nomenclature would be helpful

- In conducting tests for similarity and distinguishability
- To keep track of each patch
- To conduct group-wise, area-wise and subject-wise tests

4.4 Patch Selection

Step 1: Segmented ROI is overlaid on the original image with ROI in grayscale and the rest of the image darkened as shown in Fig. 4.3.

Step 2: Square-shaped patches of 32x32 size are selected from the ROI by selecting as many points as the required number of patches using mouse clicks (Each point representing the upper left corner of square).

Step 3 : Each patch of 32x32 size is saved in .tif format as per the nomenclature given in Section 4.3.

Step 4 : This (Step 1 to Step 3) is repeated for each malignant image.

Step 5 : (Step 2 and Step 3) is performed for each normal image.

4.5 Feature Extraction

Following are the 61 features [Ref. Table 4.1] extracted.



- 44 Gray Level Co-occurrence Matrix (GLCM) features.
- 11 Gray Level Run Length (GLRL) features .
- 6 Intensity based First order features.

TABLE 4.1: Features Extracted

| Feature No. | Feature Name | Feature No. | Feature Name |
|-------------|--------------------|-------------|--|
| 1 | Minimum value | 32 | (d=2) IDM |
| 2 | Maximum value | 33 | (d=3) IDM |
| 3 | Mean | 34 | (d=4) IDM |
| 4 | Std. Deviation | 35 | (d=1) Sum Average |
| 5 | Skewness | 36 | (d=2) Sum Average |
| 6 | Kurtosis | 37 | (d=3) Sum Average |
| 7 | (d=1) Contrast | 38 | (d=4) Sum Average |
| 8 | (d=2) Contrast | 39 | (d=1) Sum Variance |
| 9 | (d=3) Contrast | 40 | (d=2) Sum Variance |
| 10 | (d=4) Contrast | 41 | (d=3) Sum Variance |
| 11 | (d=1) Correlation | 42 | (d=4) Sum Variance |
| 12 | (d=2) Correlation | 43 | (d=1) Sum Entropy |
| 13 | (d=3) Correlation | 44 | (d=2) Sum Entropy |
| 14 | (d=4) Correlation | 45 | (d=3) Sum Entropy |
| 15 | (d=1) Energy | 46 | (d=4) Sum Entropy |
| 16 | (d=2) Energy | 47 | (d=1) Diff Entropy |
| 17 | (d=3) Energy | 48 | (d=2) Diff Entropy |
| 18 | (d=4) Energy | 49 | (d=3) Diff Entropy |
| 19 | (d=1) Homogeneity | 50 | (d=4) Diff Entropy |
| 20 | (d=2) Homogeneity | 51 | Short Run Emphasis (SRE) |
| 21 | (d=3) Homogeneity | 52 | Long Run Emphasis (LRE) |
| 22 | (d=4) Homogeneity | 53 | Low Gray-level Run Emphasis (LGRE) |
| 23 | (d=1) Entropy | 54 | High Gray-level Run Emphasis (HGRE) |
| 24 | (d=2) Entropy | 55 | Short Run Low Gray-level Emphasis (SRLGE) |
| 25 | (d=3) Entropy | 56 | Short Run High Gray-level Emphasis (SRHGE) |
| 26 | (d=4) Entropy | 57 | Long Run Low Gray-level Emphasis (LRLGE) |
| 27 | (d=1) SoS-Variance | 58 | Long Run High Gray-level Emphasis (LRHGE) |
| 28 | (d=2) SoS-Variance | 59 | Gray-level Non-Uniformity (GLN) |
| 29 | (d=3) SoS-Variance | 60 | Run Length Non-Uniformity (RLN) |
| 30 | (d=4) SoS-Variance | 61 | Run Percentage (RP) |
| 31 | (d=1) IDM | | |

4.5.1 GLCM Features

The modality of GLCM was proposed by Haralick et al. in 1970s [26]. It relies on the extraction of second order statistical features based on pixel-pair relationship[27]. The frequency of co-occurrence of various pixel-pairs in the image under consideration is tabulated to give an intuitive representation of the particular image texture. In order to extract the features GLCM is calculated for four different directions ($0^\circ, 45^\circ, 90^\circ, 135^\circ$) at four different distances ($d=1, 2, 3, 4$). Thus 16 normalized 8×8 matrices are formulated. Eleven GLCM features are extracted for each patch using each of the 16 matrices. The formulae used are given in Appendix B.

The procedure followed is given below:

- There are 480 patches. For each patch, all nine features are extracted using one GLCM.
- The above step is repeated for all the sixteen matrices. Thus sixteen sets of nine features are obtained.
- GLCM mean features are calculated by averaging the values for all four directions at a particular distance. Now, for a particular item, there will be only four different features (ie. One feature for $d=1$, one feature for $d=2$ and so on).
- This forms a set of 44 features.
- Each feature is normalised in the range $[0,1]$.

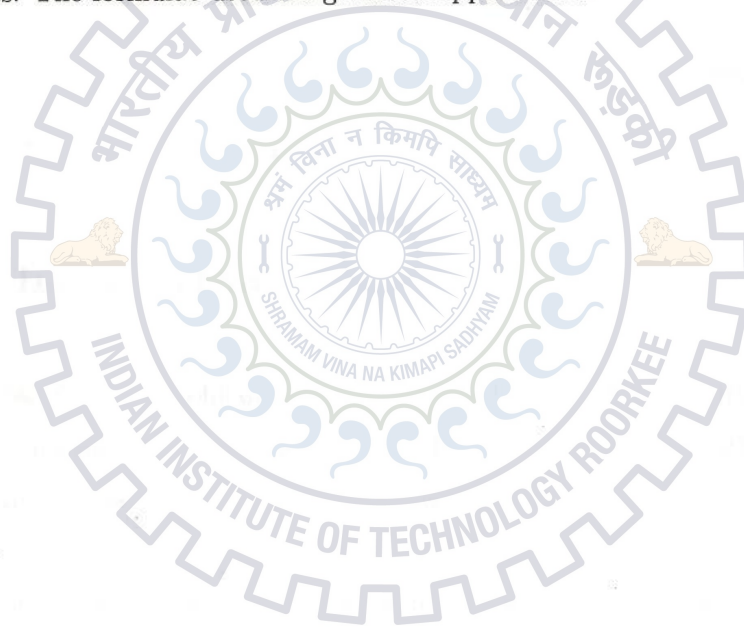
4.5.2 GLRL Features

The technique of GLRL matrix was introduced by Galloway[28] and was supplemented by Chu et al.[29] and Dasarathy and Holder[30]. Adjacent pixels with same intensity in a particular direction constitute a run and the GLRL matrix is a two dimensional tabulation of such runs against quantized pixel intensities. As evident from its definition, relatively longer runs would occur in a coarse texture while comparatively shorter runs would correspond to fine textures. The most common descriptors typically extracted from the run-length matrices are used here. In order to extract the features GLRL matrix is calculated. Using too many gray-level

intensities may cause too many null values in runlength matrix. Therefore, it is necessary to group the gray levels into smaller bins of 8 grey levels. Matrices are formed for 4 different directions ($0^\circ, 45^\circ, 90^\circ, 135^\circ$). To get a global view of texture details all the four matrices are added up. Then eleven features are calculated for each of the 480 patches. Each feature set is normalised in the range $[0,1]$. This forms a set of 11 features. The formulae[31] used are given in Appendix B.

4.5.3 Intensity Based First order Statistical Features

Features like minimum value, maximum value, mean, standard deviation, skewness and kurtosis are calculated. These six features are extracted for each of the 480 patches. Each feature set is normalised in the range $[0,1]$. This forms a set of 6 features. The formulae used are given in Appendix B.



Chapter 5

Feature Selection and Analysis

“Research is to see what everybody else has seen, and to think what nobody else has thought”

-Albert Szent-Gyrgi

5.1 Introduction

In any classification problem it is required to adopt suitable algorithm for optimal feature selection in order to improve generalisation accuracy. The other advantages of feature selection [32–34] are given below.

- Massive reduction in computation and storage requirements.
- Saving of training and computation time.
- Manoeuvring the curse of dimensionality to raise the prediction performance level.

In the current problem, two methods are used for feature selection. The algorithms and results obtained in each case are briefly discussed in this chapter. The analysis was conducted based on boxplots drawn for each of the 61 features.

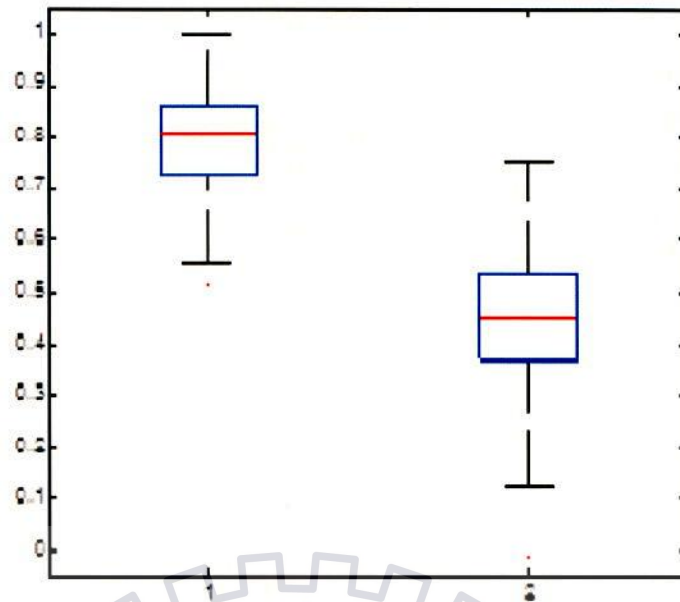


FIGURE 5.1: Sample Boxplot.

5.2 Boxplots

Box plots were introduced by John Tukey, in 1977[35] as an efficient method to display robust statistics. It is a method of visualisation of statistical measures where the data can be grouped into different classes and plotted to have a comparative analysis between the classes.

Referring to Fig. 5.1 boxplot has the following features [36] :

- The tops and bottoms of each "box" are the 25th and 75th percentiles of the samples, respectively. The distances between the tops and bottoms are the interquartile ranges.
- The line in the middle of each box is the sample median. If the median is not centered in the box, it shows sample skewness.
- The whiskers are lines extending above and below each box. Whiskers are drawn from the ends of the interquartile ranges to the furthest observations within the whisker length (the adjacent values).
- Observations beyond the whisker length are marked as outliers. By default, an outlier is a value that is more than 1.5 times the interquartile range away



FIGURE 5.2: Feature Selection – Method 1.

from the top or bottom of the box, but this value can be adjusted with additional input arguments. Outliers are displayed with a red + sign.

- Comparing box-plot medians is like a visual hypothesis test, analogous to the t-test used for means.

The position of the box in its whiskers as well as the location of the median in the box indicate the skewness of the data. A centered box and a centered median are indicative of a symmetric distribution. Whiskers which are long relative to the size of the box indicate a long tailed distribution. For a normal distribution, that is, the bell-shaped curve, the whiskers are about the same length as the box. The length of the box is a marker of the sample variability. The width of the box does not signify anything.[37]

5.3 Feature Selection – Method 1

Feature selection by method 1 is quite intuitive by virtue of which all the features which are able to distinguish between the two classes – normal and malignant – are selected by visual analysis using boxplots. Those features producing substantial overlap between their values for classes labelled different are not likely to be good at classification [Ref. Fig. 5.2]. This method is already available in the literature [38–40] and is applicable in any domain.

The algorithm is given below :

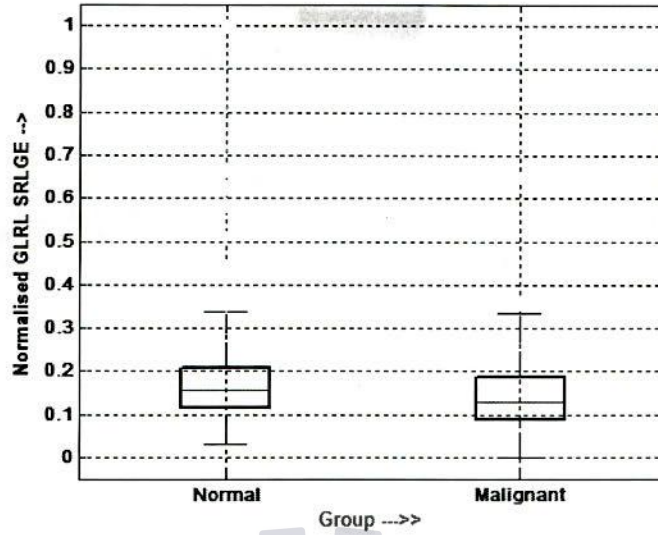


FIGURE 5.3: Boxplot of SRLGE(rejected by method 1).

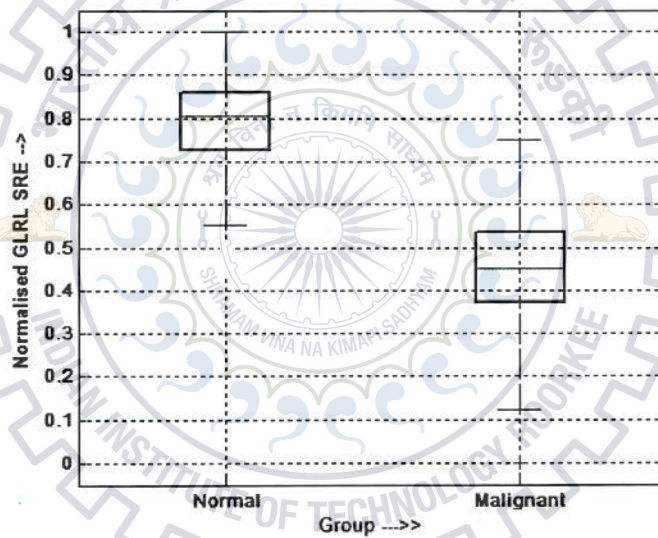


FIGURE 5.4: Boxplot of SRE(selected by method 1).

- Step 1: All the malignant and normal patches are taken
- Step 2: Boxplots are drawn for all 61 features
- Step 3: A test for distinguishability is performed between Group 1(normal) and Group 2 (malignant).
- Step 4: Check overlap among the 2 boxes in all plots. If overlap=false, corresponding features are selected while others are rejected.

TABLE 5.1: Features Selected by Method 1

| Feature No. | Feature Name | Feature No. | Feature Name |
|-------------|------------------|-------------|------------------|
| 1 | Maximum Value | 23 | SA(d=2) |
| 2 | Mean | 24 | SA(d=3) |
| 3 | SD | 25 | SA(d=4) |
| 4 | Contrast(d=2) | 26 | SV(d=1) |
| 5 | Contrast(d=3) | 27 | SV(d=2) |
| 6 | Contrast(d=4) | 28 | SV(d=3) |
| 7 | Energy(d=1) | 29 | SV(d=4) |
| 8 | Energy(d=2) | 30 | SumEntropy(d=1) |
| 9 | Energy(d=3) | 31 | SumEntropy(d=2) |
| 10 | Energy(d=4) | 32 | SumEntropy(d=3) |
| 11 | Entropy(d=1) | 33 | SumEntropy(d=4) |
| 12 | Entropy(d=2) | 34 | DiffEntropy(d=2) |
| 13 | Entropy(d=3) | 35 | DiffEntropy(d=3) |
| 14 | Entropy(d=4) | 36 | DiffEntropy(d=4) |
| 15 | sosvariance(d=1) | 37 | SRE |
| 16 | sosvariance(d=2) | 38 | LRE |
| 17 | sosvariance(d=3) | 39 | RLN |
| 18 | sosvariance(d=4) | 40 | RP |
| 19 | IDM(d=2) | 41 | GLN |
| 20 | IDM(d=3) | 42 | SRHGE |
| 21 | IDM(d=4) | 43 | LRHGE |
| 22 | SA(d=1) | 44 | LRLGE |

Finally, 44 features [Ref. Table 5.1] are selected from 61 features.

5.4 Feature Selection – Method 2

Method 2 is proposed specific to this particular context. The idea behind this method is the fact that cancer is a condition where tissues which had some regularity switch their nature to have some irregularity. Thus normality implies some kind of similarity and regularity whereas this particular regularity gives way to irregularity and differences in malignancy. In effect, the degree of variations brought about in those features is expected to be quite large where higher level of similarity was present during normal stage. So it will be a nice idea to find out initially those features with higher level of similarity in normal images. Then a test for distinguishability can be conducted on the selected features taking the entire set

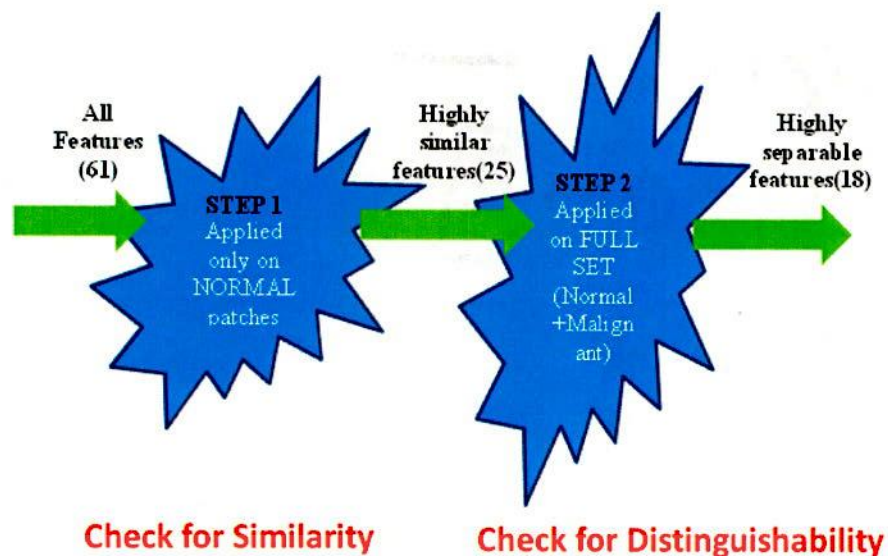




FIGURE 5.5: Feature Selection – Method 2.

of images.[Ref. Fig. 5.5]

This concept is implemented in the algorithm given below :

- Step 1: Only the normal patches (from 10 different individuals) are taken first.  
- Step 2: Boxplots are drawn for all 61 features
- Step 3: A test for similarity is performed between boxes corresponding to person 1 to 10.
- Step 4: Check overlap among the 10 boxes in all plots.If overlap=true, corresponding features are selected and all others are rejected.
- Step 5: In order to get highly correlated features among the different normal individuals, those features which are overlapping over a shorter range of normalized values are selected. A range of 0.4 is chosen after experimenting with different ranges. Thus 25 features got selected to be passed on to the next step.
- Step 6: All the malignant and normal patches are taken.
- Step 7: Boxplots are drawn for all 61 features.
- Step 8: A test for distinguishability is performed between Group 1(normal) and Group 2 (malignant).

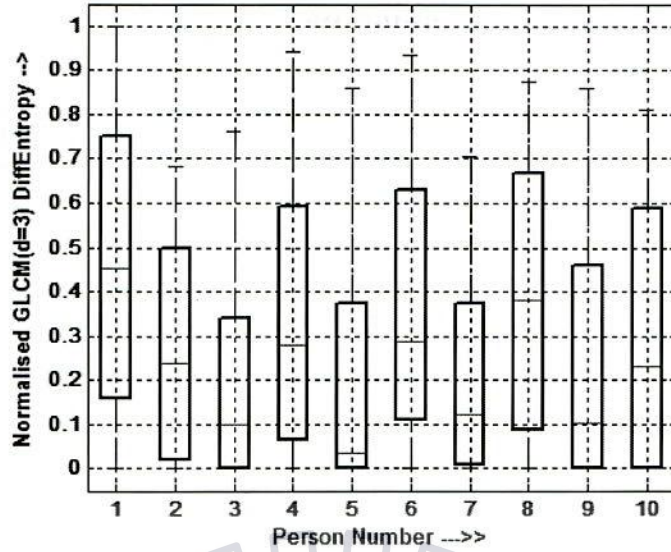


FIGURE 5.6: Boxplot of (d=3) diff. entropy (Rejected after step 5).

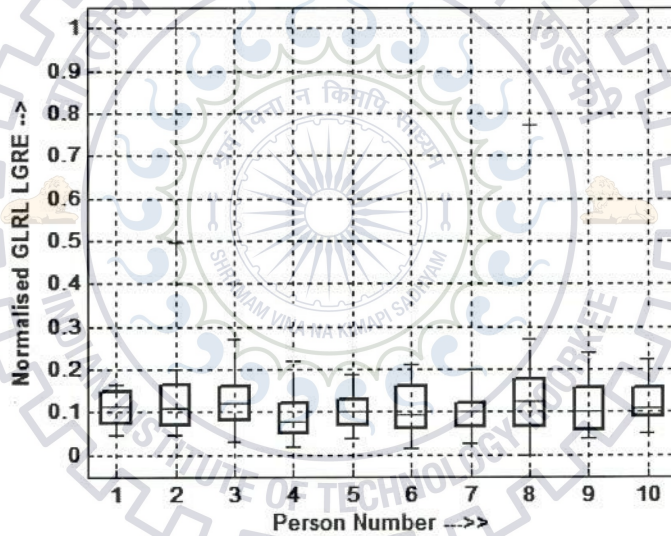


FIGURE 5.7: Boxplot of LGRE (selected after step5).

- Step 9: Check overlap among the 2 boxes in all plots. If overlap=false, corresponding features are selected and others are rejected.

Finally, 18 features [Ref. Table 5.2] were selected from 61 features.

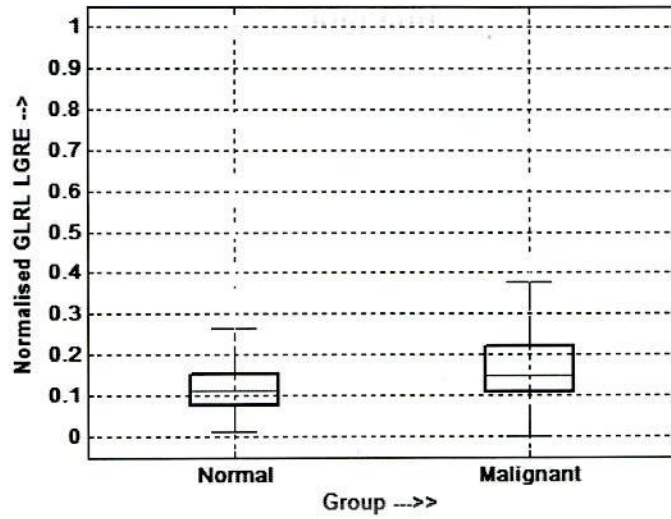


FIGURE 5.8: Boxplot of LGRE (Rejected after step 9).

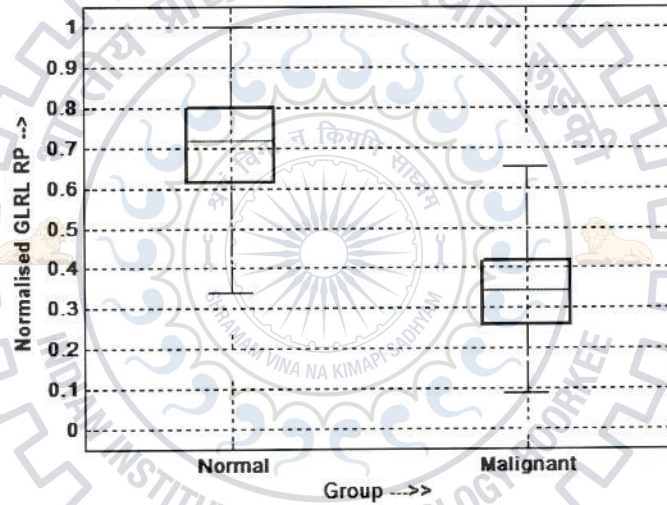


FIGURE 5.9: Boxplot of RP (selected after step 9).

TABLE 5.2: Features Selected by Method 2

| Feature No. | Feature Name | Feature No. | Feature Name |
|-------------|---------------|-------------|--------------|
| 1 | SD | 10 | SV(d=3) |
| 2 | Contrast(d=2) | 11 | SV(d=4) |
| 3 | Contrast(d=3) | 12 | SRE |
| 4 | Contrast(d=4) | 13 | LRE |
| 5 | IDM(d=2) | 14 | RLN |
| 6 | IDM(d=3) | 15 | RP |
| 7 | IDM(d=4) | 16 | SRHGE |
| 8 | SV(d=1) | 17 | LRHGE |
| 9 | SV(d=2) | 18 | LRLGE |

5.5 Conclusion

Feature selection is an indispensable part of productive data mining techniques. It reduces dimensionality of data by removing the irrelevant and the redundant features [41, 42]. In the current problem of oral lesions, feature selection is performed using boxplot analysis in two methods - (1) an intuitive Method 1 and (2) a context specific new Method 2. By application of method 1, data dimensionality is reduced from 61 to 44 whereas method 2 helps to reduce the dimensionality from 61 to 18. Now these methods need to be validated with the help of classification systems and their performances need to be analysed against standard methods of feature selection.



Chapter 6

Classification

“Many of life’s failures are people who did not realize how close they were to success when they gave up”

-Thomas Alva Edison

6.1 Introduction

In order to validate the methods of feature selection two classifiers are implemented Backpropagation based Artificial Neural Network (BPANN) and Support Vector Machines (SVM) [43].

6.2 Implementation

6.2.1 Classifier 1 BPANN

BPANN is a feed forward type neural network consisting of two steps:

1. A forward traversal of computing the net output and error at each layer.
2. A backward traversal where errors are propagated backward and weights are re-adjusted.

Learning process continues till the performance parameter is brought down below the specified goal. The iterations required for this purpose constitute the total number of epochs. In the current implementation we have used a three layered architecture – input layer, one hidden layer and output layer. The performance parameter used is Mean squared error(MSE). Levenberg - Marquardt algorithm is used to minimize MSE and train BPANN. The optimal number of neurons in the hidden layer is found out in each case by experimentation based on the least percentage of resulting error. The tests are conducted using MATLAB version 7.6. The algorithm implemented is discussed below:

Step 1 : Initialization of weights with random values

Step 2: Net output vector calculation for all input training vectors.

$$\text{net} = \Phi \sum_{i=1}^n w_i x_i \quad (6.1)$$

Step 3: Network error calculation and computation of sum squared error for all input vectors.

$$\text{Error} = \frac{1}{2} \left[\sum_{i=1}^n (D_i - \text{net}_i)^2 \right] \quad (6.2)$$

Step 4: Continue iterations till sum squared error for all training vectors is less than the specified goal.

Step 5: Calculate new weight matrix of each layer and go to step 2.

Where

Φ = Activation function (tan sigmoid for hidden layer and linear for output layer)

x = Input vector

D = desired output (target)

w = weight vector

n = number of inputs.

6.2.2 Classifier 2 SVM

SVM is one of the major approaches available for data modelling, approximation and classification. Good generalisation is achieved by finding the hyper plane which maximises the margin while bisecting the lines between the closest points on the convex hull of different datasets[44]. This is extended for nonlinear cases by mapping the data into a higher dimensional hyper plane through Kernel trick. Popular kernels are linear, RBF, polynomial and sigmoid. SVM takes structural risk minimization into account to decide the decision boundary in the provided data space. A binary SVM classifier identifies certain data points as support vectors, which best separates the data points into different classes. The advantages of SVM over other common classifiers include averting the convergence to a local minima and abstaining from overfitting the data. Thus it not only finds some means to classify the given data but makes sure that the best possible separating boundary is found out. The limitation in the use of SVM is proper selection of a kernel. In the current problem we used a RBF kernel which produced better results compared to other commonly used kernels. The experimentation is performed using LIBSVM tool[45] integrated with MATLAB 7.6. SVM generates a model depending on the training data which predicts the potential target values of the test data given only the test data attributes.

6.3 Results and Discussion

The classification is performed for 3 cases as given below. The corresponding classification matrices are shown in Table 6.1

- Case 1 – Without any feature selection
- Case 2 – After feature selection by method 1
- Case 3 – After feature selection by method 2

Training set is formulated with 120 malignant patterns and as many normal patterns. Testing set is formed with the rest of the patterns, thus containing as many

patterns as in training set. The logic behind the division is such that half of the patterns from one ROI goes into each set.

TABLE 6.1: Classification Matrices

| Classifier | Case 1 | | Case 2 | | Case 3 | |
|------------|---------------|---------------|---------------|---------------|---------------|---------------|
| BPANN | <u>TD</u> 112 | <u>FN</u> 8 | <u>TD</u> 114 | <u>FN</u> 6 | <u>TD</u> 120 | <u>FN</u> 0 |
| | <u>FA</u> 2 | <u>TN</u> 118 | <u>FA</u> 0 | <u>TN</u> 120 | <u>FA</u> 1 | <u>TN</u> 119 |
| SVM | <u>TD</u> 119 | <u>FN</u> 1 | <u>TD</u> 119 | <u>FN</u> 1 | <u>TD</u> 120 | <u>FN</u> 0 |
| | <u>FA</u> 11 | <u>TN</u> 109 | <u>FA</u> 10 | <u>TN</u> 110 | <u>FA</u> 2 | <u>TN</u> 118 |

From Table 6.1

- TD= True detection (malignant patterns classified as malignant)
- FN= False Normal (malignant patterns classified as normal)
- FA= False Alarm (normal patterns classified as malignant)
- TN= True Normal (normal patterns classified as normal)

Case 1 has 230 correct classifications and 10 misclassifications while case 2 has 234 correct classifications and 6 misclassifications with BPANN. Here case 3 has 239 correct classifications and only 1 misclassification. Using SVM, there are 228 correct classifications and 12 misclassifications with case 1 whereas case 2 correctly classified 229 and misclassified 11. Here case 3 has 238 correct classifications and only 2 misclassifications. In short misclassifications using case 3 are 1 and 2 respectively for BPANN and SVM.

TABLE 6.2: Classification accuracy based on method of feature selection

| Features Used | Classifier | Accuracy(%) |
|----------------------------------|------------|--------------|
| All 61 Features | SVM | 95.00 |
| | BPANN | 95.83 |
| 44 Features selected by method 1 | SVM | 95.42 |
| | BPANN | 97.50 |
| 18 Features selected by method 2 | SVM | 99.17 |
| | BPANN | 99.58 |

Table 6.2 displays a comparative analysis done on classification of oral cancer lesions using color images. It is quite evident that the classification performed after feature selection by method 2 has a clear upper hand over the other two cases. This work shows that a suitable method of selection of features can give better accuracy to the classification system. Feature selection by method 1 appears to be quite intuitive since all the distinguishable features are used in this case. As expected, it gives a better accuracy. But the approach used in method 2 improves the accuracy further with a reduced set of features. The reduction in number of features has positive implications in terms of computational cost, time and storage required.

6.4 Conclusion

Binary classification of oral cancer lesions in camera images is performed using texture based analysis. A dataset of 480 patterns was formulated from 20 malignant images and 51 normal images. The performance of the classification system depends on the classifier used, optimal selection of features, type of features used and several other factors. It is difficult to conclude on a classification problem based on a single system. A better approach is to design a voting system based on the accuracy of multiple classification systems.

Chapter 7

Performance Analysis

“Look for the answer inside your question”

-Jalal ad-Dn Muhammad Rumi

7.1 Introduction

The performance of the new approach of feature selection and subsequent classification proposed in the preceding chapters need to be analyzed in comparison with the already existing methods of feature selection. Standard approaches available in this field are given below:

- Principal Component Analysis (PCA)
- Sequential feature selection
- t-test
- Relief-F

In this chapter, the capability of the proposed algorithm based on boxplot analysis is tested against four standardized methods specified above. These methods bring

down the dimensionality of given data by choosing a subset of measured features which may otherwise be called predictor variables. Thus they create an optimized model for prediction. The criteria for selection may be the minimization of a particular measure of predictive error for these models fit to different subsets. Feature selection algorithms search for a subset of feature set that optimally model measured responses, depending on the required or excluded features and the size of the reduced set.

7.2 The Standard Methods

7.2.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a mathematical method which performs an orthogonal transformation and a set of observations of possibly correlated variables (complete feature set) is converted into a set of values of linearly uncorrelated variables (reduced feature set) called principal components. There will be lesser or equal number of principal components as the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables [46].

Principal component analysis is used to analyse large sized data sets, such as those coming up in data mining, chemical research, image analysis, and marketing. The factors with eigenvalues greater than 1 are considered practically important, that is, as explaining an important amount of the variability in the data, while eigenvalues less than 1 are considered practically insignificant, as explaining only a negligible portion of the data variability.

In the current problem PCA is used for reducing the feature set size from 61 to 18 and this is followed by a classification using Linear Discriminant Analysis. PCA followed by LDA is a standard approach used in classification problems since PCA helps LDA to overcome its limitations [47].

7.2.2 Sequential feature selection

A common method of feature selection is sequential feature selection. This method has two components:

An objective function, called the criterion, which the method seeks to minimize over all feasible feature subsets. Common criteria are mean squared error (for regression models) and misclassification rate (for classification models). A sequential search algorithm, which adds or removes features from a candidate subset while evaluating the criterion. Since an exhaustive comparison of the criterion value at all 2^n subsets of an n -feature data set is typically infeasible (depending on the size of n and the cost of objective calls), sequential searches move in only one direction, always growing or always shrinking the candidate feature set [48, 49]. The method has two variants:

- Sequential forward selection (SFS), where features are sequentially added to an empty candidate set until the addition of further features does not cause a reduction in the criterion.
- Sequential backward selection (SBS), where features are sequentially removed from a full candidate set until the removal of more number of features increase the criterion.

The MATLAB 7.6 Statistics Toolbox function 'sequentialfs' is used for sequential feature selection. Input arguments include predictor and response data and a function handle to a file implementing the criterion function. Optional inputs allows to specify SFS or SBS, required or excluded features, and the size of the feature subset.

7.2.3 t-Test

Absolute value two-sample t-Test with pooled variance estimate is used as the criterion to analyse the suitability of each feature in classifying two labelled classes. This criterion is passed as a parameter in the MATLAB command 'rankfeatures' and the features are ranked accordingly. Another argument is the number of features to be selected. In the current problem first 18 features are selected by this method and it is followed up by a SVM classification [50].

7.2.4 Relief-F

The independence of attributes has assumed by heuristic measures for estimating the quality of attributes such that, their performance is poor in domains with strong dependencies between attributes. Relief and its extension ReliefF are capable of correctly estimating the quality of attributes in classification problems with strong dependencies between attributes. They provide a global view by exploiting local information obtained from different contexts. Regression ReliefF and ReliefF provide a unified view on estimating the attribute quality in regression and classification [51].

Relief-F is a feature selection methodology that selects instances randomly, and change the scores of the feature relevance on the basis of nearest neighbor. It is one of the most successful strategies in feature selection due to the merits it offer. In MATLAB 7.6, `relieff(X,Y,K)` computes ranks and weights of attributes (predictors) for input data matrix X and response vector Y using the ReliefF algorithm for classification. For classification, it uses K nearest neighbors per class. There are two output variables: RANKED and WEIGHT. RANKED gives indices of columns in X in the order of attribute importance. That is, RANKED(1) is the index of the most important predictor. WEIGHT gives attribute weights in a range of -1 to 1. Here, large positive weights are assigned to significant attributes [52].

Attribute weights and ranks computed by `relieff` usually depend on K. If K is set to 1, the estimates computed by `relieff` can be unreliable for noisy data. If you set K

to a value comparable with the number of observations (rows) in X , relief may not find significant attributes. A better idea is to start with $K = 10$ and investigate the stability and reliability of relief ranks and weights for various values of K [53].

7.3 Performance Parameters

Performance of the two cases of classification completed in Chapter 6 is compared with that of four standard cases specified in the previous section with the help of following commonly used parameters [54].

- Sensitivity
- Specificity
- Accuracy

These are popular in quantifying the classification capability of different models.

7.3.1 Sensitivity

Sensitivity of the classifier model is the proportion of true positives which are correctly classified by the model. Thus it quantifies the ability of the model to be sensitive to positive results. In our problem, true detection of malignant cases is considered as true positive. Therefore sensitivity is defined [55] as

$$\text{Sensitivity} = \frac{\text{Number of True Detections}}{\text{Total Number of Malignant Cases}}$$

A model with high sensitivity is regarded as a reliable indicator when the result is negative, as it rarely misses true detections among those who are actually cancerous. Thus, a negative result (ie. a case of declaring normal) from a test with high specificity means a high probability of the absence of disease.

7.3.2 Specificity

Specificity of the classifier model is the proportion of true negatives which are correctly classified by the model. Thus it quantifies the ability of the model to be specific while identifying negative results. In our problem, true normal detection is considered as true positive. Therefore specificity is defined as

$$\text{Specificity} = \frac{\text{Number of True Normals}}{\text{Total Number of Normal Cases}}$$

A model with high specificity regarded as a reliable indicator when the result is positive, as it rarely misses out on finding true normals. Thus, a positive result (ie. a case of detection of cancer) from a test with high specificity means a high probability of the presence of disease.

7.3.3 Accuracy

In binary classification, accuracy of the classifier model is the most important parameter for analysing its performance. It is the proportion of true results (both true detections and true normals) in the population. An accuracy of 100% implies that the measured values are exactly the same as the values used for training. Therefore accuracy is defined as

$$\text{Accuracy} = \frac{\text{Number of True Detections} + \text{Number of True Normals}}{\text{Total Number of Cases}}$$

7.4 Results and Discussion

7.4.1 PCA followed by LDA

Principal Component Analysis (PCA) is used to reduce the dimensionality from 61 features to 18 features and classification is performed using Linear Discriminant Analysis (LDA). Classification Matrix is given in Fig 7.1

| | |
|-----------------------------------|---------------------------------|
| <u>True Detection (TD)</u> 114 | <u>False Normals (FN)</u> 6 |
| <u>False Alarms (FA)</u> 7 | <u>True Normals (TN)</u> 113 |

FIGURE 7.1: Classification Matrix : PCA + LDA

Result 1 : Sensitivity = 0.9475

Result 2 : Specificity = 0.9442

Result 3 : Accuracy = 94.58 %

7.4.2 SFS followed by SVM

TABLE 7.1: Features Selected by SFS

| Feature No. | Feature Name | Feature No. | Feature Name |
|-------------|------------------|-------------|------------------|
| 1 | Minimum | 13 | SV(d=2) |
| 2 | Kurtosis | 14 | SV(d=3) |
| 3 | Contrast(d=1) | 15 | SV(d=4) |
| 4 | Contrast(d=2) | 16 | SumEntropy(d=4) |
| 5 | Contrast(d=4) | 17 | DiffEntropy(d=2) |
| 6 | Entropy(d=2) | 18 | SRE |
| 7 | sosvariance(d=1) | 19 | RP |
| 8 | sosvariance(d=2) | 20 | GLN |
| 9 | IDM(d=1) | 21 | LGRE |
| 10 | IDM(d=3) | 22 | SRHGE |
| 11 | IDM(d=4) | 23 | LRLGE |
| 12 | SV(d=1) | | |

Sequential forward selection (SFS) is used to reduce the dimensionality from 61 features to 23 features and classification is performed using Support Vector Machines (SVM). Features selected are given in Table 7.1. Classification Matrix is given in Fig 7.2

| | |
|-----------------------------------|---------------------------------|
| <u>True Detection (TD)</u> 120 | <u>False Normals (FN)</u> 0 |
| <u>False Alarms (FA)</u> 12 | <u>True Normals (TN)</u> 108 |

FIGURE 7.2: Classification Matrix : SFS + SVM

Result 1 : Sensitivity = 1.0

Result 2 : Specificity = 0.9

Result 3 : Accuracy = 95 %

7.4.3 t-Test followed by SVM

Absolute value two-sample t-test with pooled variance estimate is used to reduce the dimensionality from 61 features to 18 features and classification is performed using Support Vector Machines (SVM). Features selected are given in Table 7.2. They are ranked according to their scores in the same order. Classification Matrix is given in Fig 7.3

| | |
|-----------------------------------|---------------------------------|
| <u>True Detection (TD)</u> 120 | <u>False Normals (FN)</u> 0 |
| <u>False Alarms (FA)</u> 10 | <u>True Normals (TN)</u> 110 |

FIGURE 7.3: Classification Matrix : t-Test + SVM

Result 1 : Sensitivity = 1.0

Result 2 : Specificity = 0.9167

Result 3 : Accuracy = 95.83 %

TABLE 7.2: Features Selected by t-Test

| Feature No. | Feature Name | Feature No. | Feature Name |
|-------------|-----------------|-------------|-----------------|
| 1 | SRE | 10 | Entropy(d=4) |
| 2 | RLN | 11 | Entropy(d=3) |
| 3 | GLN | 12 | Energy(d=4) |
| 4 | RP | 13 | SumEntropy(d=1) |
| 5 | SD | 14 | Energy(d=3) |
| 6 | Maximum | 15 | Entropy(d=2) |
| 7 | SumEntropy(d=4) | 16 | Energy(d=2) |
| 8 | SumEntropy(d=3) | 17 | Entropy(d=1) |
| 9 | SumEntropy(d=2) | 18 | Energy(d=1) |

7.4.4 ReliefF followed by SVM

ReliefF algorithm is used to reduce the dimensionality from 61 features to 18 features and classification is performed using Support Vector Machines (SVM). Features selected are given in Table 7.3. Classification Matrix is given in Fig 7.4

| | |
|-----------------------------------|---------------------------------|
| <u>True Detection (TD)</u> 119 | <u>False Normals (FN)</u> 1 |
| <u>False Alarms (FA)</u> 4 | <u>True Normals (TN)</u> 116 |

FIGURE 7.4: Classification Matrix : ReliefF + SVM

Result 1 : Sensitivity = 0.9916

Result 2 : Specificity = 0.9667

Result 3 : Accuracy = 97.91 %

TABLE 7.3: Features Selected by ReliefF

| Feature No. | Feature Name | Feature No. | Feature Name |
|-------------|--------------|-------------|------------------|
| 1 | SRE | 10 | Kurtosis |
| 2 | RP | 11 | HGRE |
| 3 | SRLGE | 12 | Maximum |
| 4 | RLN | 13 | LRE |
| 5 | LRLGE | 14 | Skewness |
| 6 | GLN | 15 | DiffEntropy(d=1) |
| 7 | SRHGE | 16 | IDM(d=1) |
| 8 | LGRE | 17 | Energy(d=4) |
| 9 | SD | 18 | Energy(d=1) |

7.5 Conclusion

Table 7.4 shows the summary of final results obtained by performance analysis. Here three performance parameters viz. sensitivity, specificity, accuracy are tabulated for classification of oral lesions performed in eight different ways. Classification cases 7 and 8 are performed after the proposed method (Method 2) of feature selection as given in Chapter 5. The performance of these cases are compared with different standard approaches. The standard approaches are :

1. Feature selection by PCA followed by classification by LDA.
2. Feature selection by SFS followed by classification by SVM.
3. Feature selection by t-Test followed by classification by SVM.
4. Feature selection by ReliefF followed by classification by SVM.
5. Feature selection by Method 1 followed by classification by SVM (Ref.Chap5).
6. Feature selection by Method 1 followed by classification by BPANN.

TABLE 7.4: Performance Analysis Table

| Sl.No. | Method | TD | FN | FA | TN | Sensitivity | Specificity | Accuracy |
|--------|-----------------|-----|----|----|-----|-------------|-------------|----------|
| 1 | PCA + LDA | 114 | 6 | 7 | 113 | 0.9475 | 0.9442 | 94.58 |
| 2 | SFS + SVM | 120 | 0 | 12 | 108 | 1 | 0.9 | 95 |
| 3 | t-Test + SVM | 120 | 0 | 10 | 110 | 1 | 0.9167 | 95.83 |
| 4 | ReliefF + SVM | 119 | 1 | 4 | 116 | 0.9916 | 0.9667 | 97.91 |
| 5 | Method 1+SVM | 119 | 1 | 10 | 110 | 0.9916 | 0.9167 | 95.42 |
| 6 | Method 1+BPANN | 114 | 6 | 0 | 120 | 0.95 | 1 | 97.5 |
| 7 | Method 2+SVM | 120 | 0 | 2 | 118 | 1 | 0.9833 | 99.58 |
| 8 | Method 2+ BPANN | 120 | 0 | 1 | 119 | 1 | 0.9916 | 99.17 |

The observations derived are :

- Accuracy of classification models preceded by the proposed method of feature selection is observed to be the best among all the cases with SVM giving 99.17 % and BPANN giving a whopping 99.58 % .
- Sensitivity and specificity of the classification models preceded by the proposed method of feature selection is observed to be the best cases.
- The potential number of features to be selected should be specified in case of PCA,t-Test and ReliefF. This is not an easy task since the number of relevant features could not be predicted without any analysis. This trouble is manoeuvred in case of Method 1,Method 2 and SFS where all features are analysed and relevant features are chosen.

From the observations, it is concluded that the classifier models preceded by the new method of feature selection outperforms all the standard approaches by quite convincing margin. Thus the classifiers (Method 2 + SVM) and (Method 2 + BPANN) are chosen to be the best for the problem of oral lesion classification.

Chapter 8

Multi-class Classification of Malignant Lesions Using ANN

*“Sure he that made us with such large discourse,
Looking before and after, gave us not
That capability and god-like reason
To rust in us unus’d”*

-William Shakespeare

8.1 Introduction

Texture based features derived from Grey Level Co-occurrence Matrix (GLCM) and Grey Level Run-Length (GLRL) matrix are widely used for image characterization. Once the objective of obtaining the binary classification between normal and malignant patterns was successfully completed as in preceding chapters, the subsequent requirement specified by the doctor is to have a multi-class classification within the set of available malignant cases. In this chapter, suitably selected texture discriminating features are used for classification of oral cancer lesions in digital camera images into six groups. Backpropagation based Artificial Neural Network (BPANN) is used to compare and validate the performance of different

feature sets. The classification accuracy is observed to improve with combination of GLCM, GLRL and intensity based first order features. Further improvement in accuracy is obtained by application of feature selection using boxplot analysis.

8.2 Image Database and methodology

The material analysed in the current work was recorded under customized conditions at Himalayan Institute of Medical Sciences, Jolly Grant, Dehradun , India. The digital true color images were recorded using Sony cybershot digital still picture camera (4x optical zoom, 10.1MP) with standard settings. The clinical diagnoses have been verified via histopathology. The database comprises of 16 images of malignant oral lesions. As an initial step towards feature analysis and classification, the given set of malignant images is identified to be from following 6 groups [Ref. Table 8.1]

TABLE 8.1: Six Groups of Malignant Cases

| Group No | Diagnosis |
|----------|---|
| 1 | Carcinoma Retromolar area |
| 2 | Carcinoma angle of mouth – ulcero-proliferative |
| 3 | Verrucous carcinoma buccal mucosa |
| 4 | Carcinoma Buccal mucosa – Ulcero-proliferative |
| 5 | Carcinoma Tongue – lateral border of tongue |
| 6 | Carcinoma Palate – Ulcero-proliferative growth |

Images acquired by a standard digital camera are passed through the computer aided diagnostic tool for binary classification of oral cancer lesions. Those images whose patterns are identified and classified as malignant are given to a segmentation program which identifies the ROI. Patches of 32x32 size are extracted from each of the malignant ROI. 61 features based on texture (GLCM), run-length (GLRL) and intensity variation are extracted. Suitable features are selected in

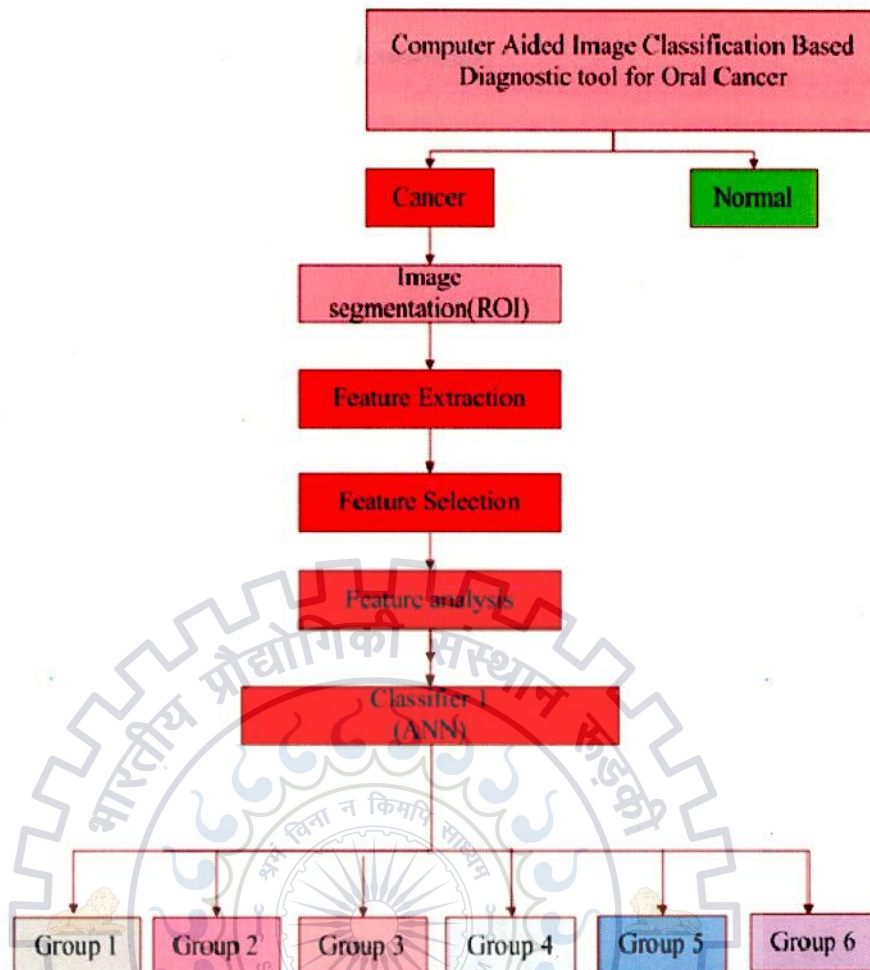


FIGURE 8.1: Proposed Methodology.

two different methods with the help of boxplots. Finally classification is performed using Backpropagation Based Artificial Neural Network. The behaviour of classifier towards the different sets of features is analysed. This classifier will serve as the building stone for the potential oral cancer prognostic tool. The Outline of the proposed methodology for multi-class classification of malignant images into six groups is given in Fig.8.1.

8.3 Segmentation of ROI

The purpose here is to extract the lesion area from the rest of the image which is constituted by the normal (healthy) tissues. The method of active contour without edges is implemented due to following reasons.

- It is a semi-automatic technique.
- It reduces the delineation time.
- It works well with all the test samples despite the high degree of variability involved.
- It is based on curve evolution and keeps on finding the boundary even though it is not well-defined by gradient.

The algorithm implemented is already explained in Section 4.2.

8.4 Feature Extraction

Segmented ROI is overlaid on the original image. 192 Patches of 32x32 size are selected and saved in .tif format. A set of 61 features were extracted as discussed below:

1. Gray Level Co-occurrence Matrix (GLCM) features (44).
2. Gray Level Run Length (GLRL) features (11).
3. Intensity based First order features (6).

8.4.1 GLCM Features

GLCM is calculated for four different directions ($0^\circ, 45^\circ, 90^\circ, 135^\circ$) at four different distances ($d=1,2,3,4$). Thus 16 normalized 8x8 matrices are formulated. Features extracted for each patch using each of the 16 matrices include Contrast, Correlation, Energy (Angular Second Moment), Homogeneity, Entropy, Sum of squares variance, Inverse Difference Moment (IDM), Sum Average (SA), Sum Variance (SV), Sum Entropy (SE) and Difference Entropy (DE).

The procedure followed is given below:

- There are 480 patches. For each patch, all nine features are extracted using one GLCM.

- The above step is repeated for all the sixteen matrices. Thus sixteen sets of nine features are obtained.
- GLCM mean features are calculated by averaging the values for all four directions at a particular distance. Now, for a particular item, there will be only four different features (ie. One feature for $d=1$, one feature for $d=2$ and so on).
- This forms a set of 44 features.
- Each feature is normalised in the range $[0,1]$.

The formulae used are given in Appendix B.

8.4.2 GLRL Features

GLRL matrix is calculated for four different orientations ($0^\circ, 45^\circ, 90^\circ, 135^\circ$) and are added up to get a global look of texture variations. Features extracted for each patch include Short Run Emphasis (SRE), Long Run Emphasis (LRE), Low Gray-level Run Emphasis (LGRE), High Gray-level Run Emphasis (HGRE), Short Run Low Gray-level Emphasis (SRLGE), Short Run High Gray-level Emphasis (SRHGE), Long Run Low Gray-level Emphasis (LRLGE), Long Run High Gray-level Emphasis (LRHGE), Gray-level Non-Uniformity, Run Length Non-Uniformity and Run Percentage (RP). Thus 11 features are extracted for each of the 192 patterns. Each feature is normalised in the range $[0,1]$. The formulae used are given in Appendix B.

8.4.3 Intensity Based First order Statistical Features

Features like minimum value, maximum value, mean, standard deviation, skewness and kurtosis are calculated. Thus six features are extracted for each of the 192 patterns. Each feature is normalised in the range $[0,1]$. This forms a set of 6 features. The formulae used are given in Appendix B.

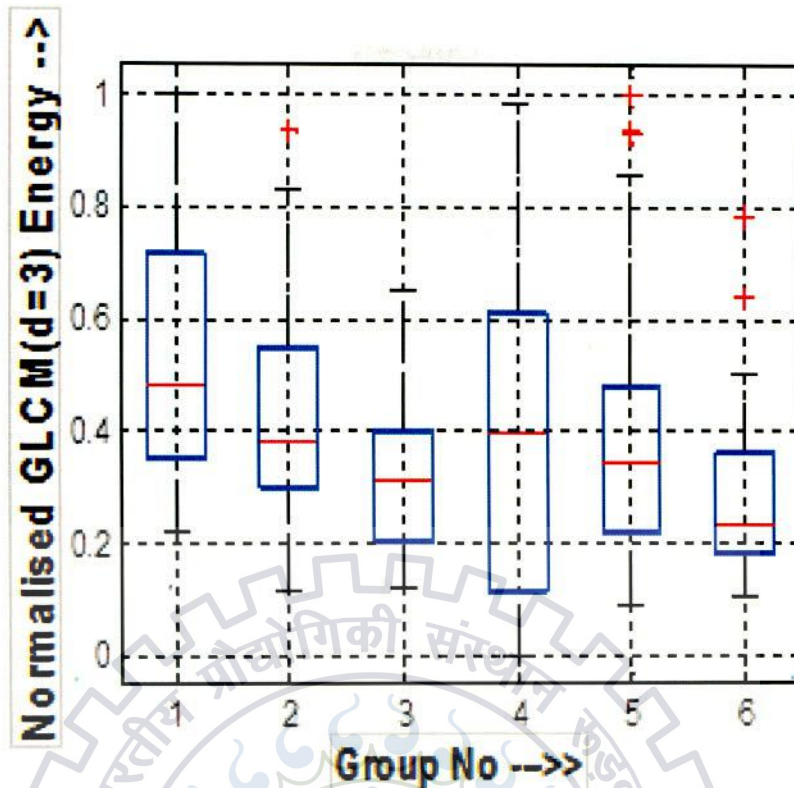


FIGURE 8.2: Boxplot of GLCM Energy ($d=3$). This feature is rejected due to overlap.

8.5 Feature Analysis and selection

In order to handle the problems associated with the curse of dimensionality, it is a common practice in any classification problem to adopt a suitable method for optimal feature selection. The level of sophistication adapted during this phase helps in improving generalisation accuracy while reducing the computation time and storage requirements. Here a method of exploring the distinguishable features with the help of boxplots is conceived. Boxplots are created for all 61 normalized features with the respective feature data along Y-axis and the groups along the X-axis. With the help of boxplots, all the 61 features are studied to analyse and decide the features which are suitable to distinguish between the groups taking two groups at a time. Thus 15 group-pairs are formed.

$$\begin{pmatrix} 6 \\ 2 \end{pmatrix} = 15$$

Now, Table 8.2 is formulated where all the features and their corresponding distinguishabilities are shown. Distinguishability of a particular feature is indicated by '1' whereas overlap is indicated by '0'.



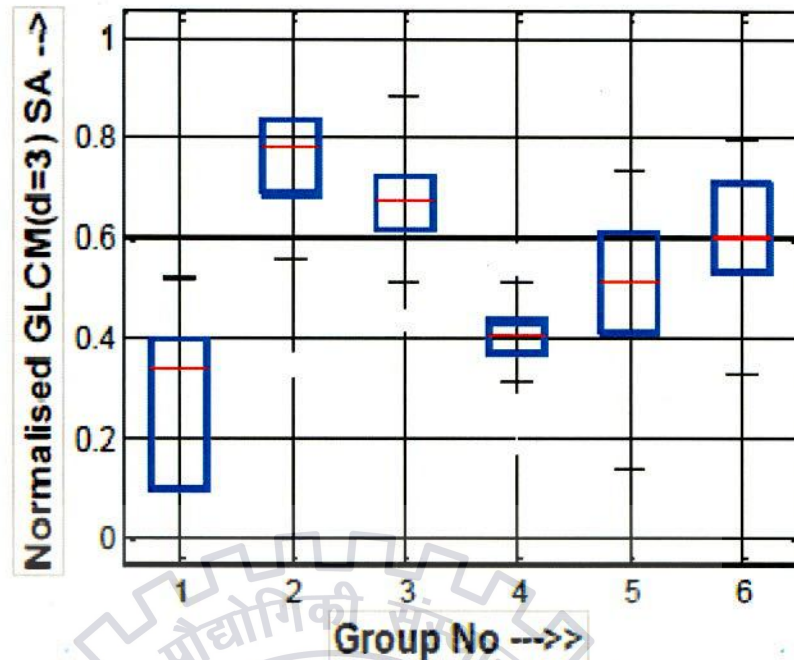


FIGURE 8.3: Boxplot of GLCM SA ($d=3$). This feature is selected owing to distinguishability.

Feature selection is performed by comparing the overlapping of box area created by different groups in a plot for a particular feature. Those features which are useful in distinguishing between at least 10 out of 15 pairs of groups are selected. Finally, 11 features are selected. They are GLCM Sum Average ($d=1,2,3,4$), Sum of squares variance ($d=1,2,3,4$), First order Intensity mean, SRE and HGRE.

8.6 Experimental Results

In order to validate the feature selection method and to find out the optimal set of features, the patterns in the dataset are applied to a classification system designed using Back Propagation based Artificial Neural Network (BPANN). BPANN is a feed forward type neural network consisting of two steps:

1. A forward traversal of computing the net output and error at each layer.
2. A backward traversal where errors are propagated backward and weights are re-adjusted.

Learning process continues till the performance parameter is brought down below the specified goal. The iterations required for this purpose constitute the total number of epochs [56, 57]. In the current implementation we have used a three layered architecture – input layer, one hidden layer and output layer. The performance parameter used is Mean squared error(MSE). Levenberg - Marquardt algorithm is used to minimize MSE and train BPANN. The optimal number of neurons in the hidden layer is found out in each case by experimentation based on the least percentage of resulting error. The tests are conducted using MATLAB version 7.6. The algorithm implemented is discussed below:

Step 1 : Initialization of weights with random values

Step 2: Net output vector calculation for all input training vectors.

$$\text{net} = \Phi \sum_{i=1}^n w_i x_i \quad (8.1)$$

Step 3: Network error calculation and computation of sum squared error for all input vectors.

$$\text{Error} = \frac{1}{2} \left[\sum_{i=1}^n (D_i - \text{net}_i)^2 \right] \quad (8.2)$$

Step 4: Continue iterations till sum squared error for all training vectors is less than the specified goal.

Step 5: Calculate new weight matrix of each layer and go to step 2.

Where

Φ = Activation function (tan sigmoid for hidden layer and linear for output layer)

x = Input vector

D = desired output (target)

w = weight vector

n = number of inputs.

Training set is formulated with 75 percent of patterns taken from each group of malignancy. Testing set is formed with the rest of the patterns , thus containing 25

percent of total number of patterns from each malignant class. The classification is performed for four cases as given below.

Case 1: Using all GLCM features

Case 2: Using all GLRL features

Case 3: Using All 61 features

Case 4: Using 11 features selected by boxplot analysis

Classification results are shown in Table 8.3. Optimal number of hidden layer neurons required in all the cases are also shown. It is clear that classification accuracy improves as the features derived from GLCM and GLRL are combined. Accuracy increases further with the introduction of feature selection method into the system. The classification matrix for the best case (i.e. case 4) is shown in Table 8.4.

TABLE 8.3: Six-class Classification Results

| Case | Performance | | |
|------|--------------|----------------------|--------------|
| | Features | Hidden layer neurons | Accuracy (%) |
| 1 | 44 (GLCM) | 36 | 89.58 |
| 2 | 11 (GLRL) | 31 | 85.41 |
| 3 | 61(All) | 16 | 91.66 |
| 4 | 11(selected) | 25 | 97.92 |

8.7 Conclusion

Multiclass classification of oral cancer lesions in color images is performed using different feature sets with the help of Backpropagation based ANN classifier. A set of 61 features is formulated and applied on 192 sections of images taken from 16 patients. Grouping is done into six groups as described by the doctor. It is

TABLE 8.4: Classification Matrix for case 4

| | Ground Truth Class (Assigned by doctor) | | | | | | |
|------------------------------------|---|--------------|-----|-----|-----|-----|------|
| | Groups | 1 | 2 | 3 | 4 | 5 | 6 |
| Class Predicted | 1 | 6 | 0 | 0 | 0 | 0 | 0 |
| | 2 | 0 | 6 | 0 | 0 | 0 | 0 |
| | 3 | 0 | 0 | 9 | 0 | 0 | 1 |
| | 4 | 0 | 0 | 0 | 9 | 0 | 0 |
| | 5 | 0 | 0 | 0 | 0 | 12 | 0 |
| | 6 | 0 | 0 | 0 | 0 | 0 | 5 |
| Individual Class(%) Accuracy | | 100 | 100 | 100 | 100 | 100 | 83.3 |
| Overall Classification Accuracy(%) | | 97.92 | | | | | |

observed that a combination of GLCM and GLRL features is instrumental in improving the classification accuracy. The classification performed after selecting the optimal features using boxplot analysis turns out to be the best case. Remarkable reduction in computation time is also observed in case of classification using reduced set of features compared to other cases. The efficiency of the proposed system offers a vindication of the relevance of texture features in discriminating oral cancer lesions. Such a classification of malignancies is helpful in prognosis and treatment of oral cancer which is the most common form of cancer in India.

Chapter 9

Conclusion and Future Scope

“Somewhere, something incredible is waiting to be known”

-Carl Sagan

This chapter aims to narrate the summary of the work done in this thesis and concludes by precisely enlisting the various targets attained. The musings about the possible further developments of the implemented system are also discussed towards the end.

9.1 Conclusion

Chapter 1 brings out the relevance of the proposed work. The literature review conducted in Chapter 2, facilitates analysis of the various systems for oral cancer classification. The importance of various steps involved in a classification system are identified. A methodology is proposed and the image database is outlined in Chapter 3.

The major target of this thesis work is to achieve binary classification of images into cancerous and noncancerous. This elementary goal is reached using various steps as described over the Chapter 4 to Chapter 6. Chapter 4 explains the active contour method used for image segmentation and introduces the various features

extracted for the purpose. In Chapter 5, the various methods of feature selection are described. The method 2 algorithm developed in this chapter is a brand new approach and is formulated based on valuable suggestions from medical domain experts. The mathematical implementation is done with the help of boxplots. In Chapter 6 two popular classifiers - BPANN and SVM - are implemented based on different methods of feature selection. Here it is convincingly proved that feature selection by method 2 provides better accuracy to the classification system. In Chapter 7 the performance of the developed method is further analysed in comparison with four standard methods of feature selection. The performance parameters used are sensitivity, specificity and accuracy of the SVM classification performed after each of the feature selection methods. The clear-cut advantage of the proposed method over the existing methods is brought out explicitly with the help of the performance analysis table [Ref. Table 7.4]. This concludes the comprehensive achievement of the primary target.

The supplementary work carried out includes a multi-class classification of malignant cases. Available malignant cases are classified into six groups using ANN classifier. Classification is done in four cases by varying the set of features used in each case. The best performance based on accuracy (97.92 %) is observed with feature set containing 11 features selected by boxplot analysis. The thesis work is concluded with the development of an application environment for oral cancer diagnosis designed on the grounds of the best method selected .

In a concise format, the notable contributions made by this work can be enlisted as given below:

- A context specific method of feature selection is proposed, implemented and validated for the purpose of binary classification of oral cavity images into cancerous and non-cancerous.
- The novel approach is developed into a Computer Aided Diagnostic (CADx) Tool so that it can be used in mass screening initiatives.
- The Graphical User Interface (GUI) developed [Ref. Fig. 9.1] accepts a camera image as input and outputs an inference on the nature of the image

based on analysis and classification performed on the spot. Based on the inference, one can decide whether to go for a biopsy or not.

- GUI is developed based on the new method - Feature selection in Method 2 followed by SVM classification[Ref. Fig. 9.2].
- A six-class classification of malignant images is performed using ANN classifier.



FIGURE 9.1: Welcome screen of GUI

9.2 Future Scope

- More number of images may be added to the database and the system performance will keep on increasing with the addition of images.
- Multiple classifier models may be developed . A voting system designed based on the accuracy and other performance parameters of these models will add to the reliability and robustness of the CADx Tool.
- Addition of rotation and scale invariant features (eg: Gabor wavelet features) to the current feature set can compensate for the possible discrepancies that could creep into the image analysis system due to sloppy acquisition.

- Automation of the algorithm designed for feature selection can become utile when it comes to modification and configuration of the system.
- The system configuration utilities can also be incorporated into the CADx tool.
- The tool developed can be used in mass screening initiatives with required hardware including a common digital camera and a laptop. Such a tool will act as a supplement to the existing clinical methods and reduce the burden on the medical practitioner in two respects.
 1. Any volunteer taking part in a screening programme can use this device to predict within seconds whether the patient has to go for a biopsy and further treatment.
 2. It can also be used for quantitative assessment during follow-up to the treatment by comparative studies.
- In practice, this idea will save about Rs.10 lakh estimated cost of biopsy for a village in rural India (at a rate of Rs.1000 per biopsy with 1000 people attending) per screening program. Moreover it will save some unfortunate victims from unaccountable pain and sufferings.

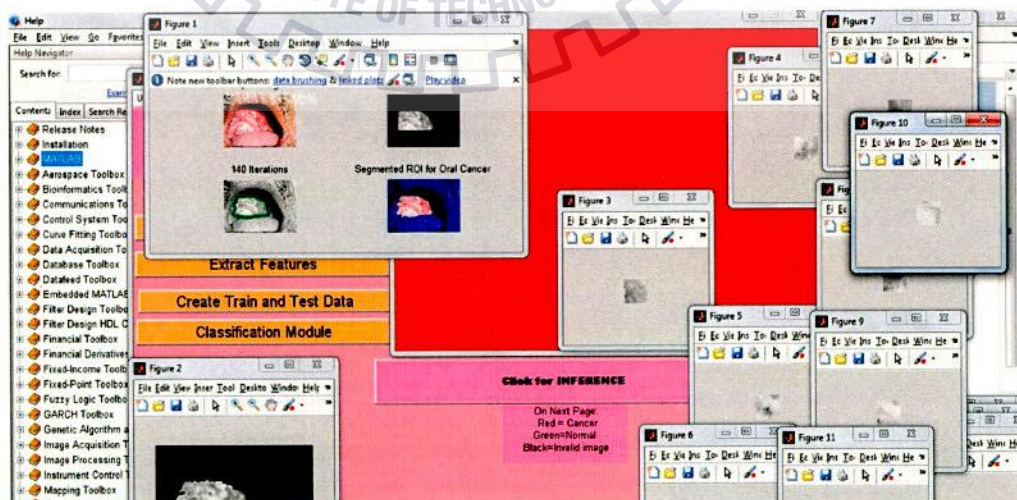


FIGURE 9.2: GUI in action

9.3 Awards and Recognition

The idea of oral cancer diagnostic tool fetched a place in the finals of National Level competition for Anjani Mashelkar Inclusive Innovation Award 2012 and received a letter of appreciation [Fig.9.3] from the foundation. The theme of inclusive innovation is to solve problems that NEED to be solved as opposed to those that CAN be solved.

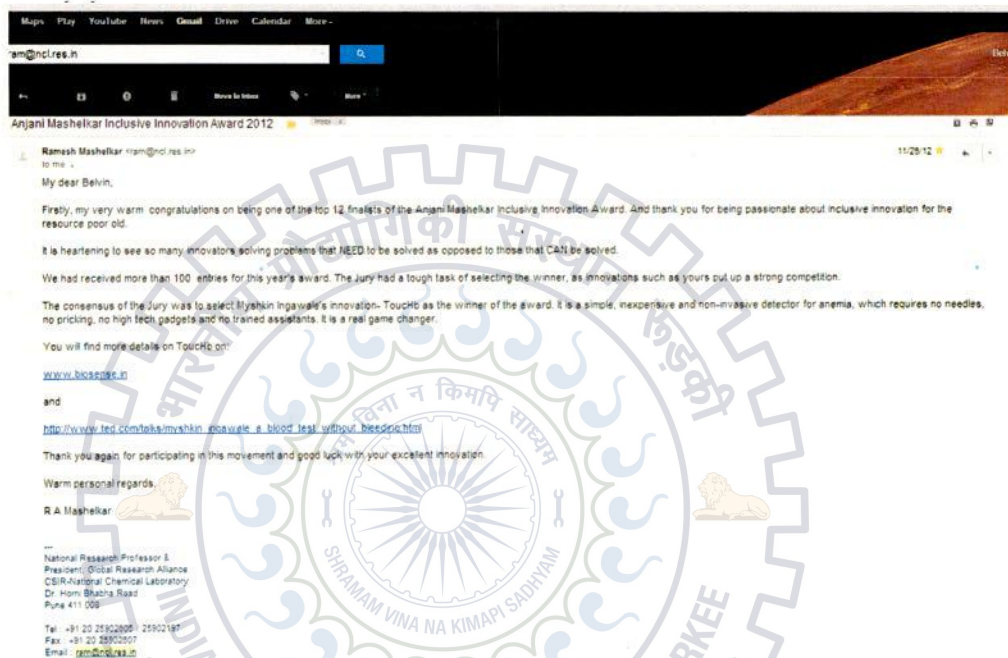


FIGURE 9.3: Screenshot of Letter of Appreciation

Publications

1. B. Thomas, V. Kumar and S. Saini ,”Identification and Classification of Oral Cancer Lesions in Color Images”, Submitted to *Springer Journal of IE(I)-Signals and Communications Series*
2. B. Thomas, V. Kumar and S. Saini ,”Texture Analysis Based Segmentation and Classification of Oral Cancer Lesions in Color Images Using ANN”, Submitted to *2013 IEEE International Conference on Signal Processing, Computing and Control*, Shimla, India sponsored by IEEE Communications Society (Expected date of acceptance : 15 June 2013)



References

- [1] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Reading, MA: Addison-Wesley, 1992.
- [2] C. Scully, J. Bagan, C. Hopper, and J. Epstein, "Oral cancer: current and future diagnostic techniques." *Am J Dent*, vol. 21, no. 4, pp. 199–209, 2008.
- [3] D. Rizzolo, C. Hanifin, and T. A. Chiodo, "Oral cancer: how to find this hidden killer in 2 minutes," *JAAPA : official journal of the American Academy of Physician Assistants*, vol. 10, no. 29, pp. 42–47, 2008. [Online]. Available: http://journals.lww.com/jaapa/Fulltext/2007/10000/Oral_cancer_How_to_find_this_hidden_killer_in_2.21.aspx
- [4] C. AY, D. C, and J. A, "US mortality rates for oral cavity and pharyngeal cancer by educational attainment," *Archives of OtolaryngologyHead and Neck Surgery*, vol. 137, no. 11, pp. 1094–1099, 2011. [Online]. Available: <http://dx.doi.org/10.1001/archoto.2011.180>
- [5] Y.-Y. Wang, S.-C. Chang, L.-W. Wu, S.-T. Tsai, and Y.-N. Sun, "A color-based approach for automated segmentation in tumor tissue classification," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, 2007, pp. 6576–6579.
- [6] A. Chodorowski, U. Mattsson, and T. Gustavsson, "Oral lesion classification using true-color images," pp. 1127–1138, 1999. [Online]. Available: [+http://dx.doi.org/10.1117/12.348507](http://dx.doi.org/10.1117/12.348507)
- [7] Y.-Y. Wang, S.-C. Chang, L.-W. Wu, S.-T. Tsai, and Y.-N. Sun, "A color-based approach for automated segmentation in tumor tissue classification," in

- Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, 2007*, pp. 6576–6579.
- [8] W. Jung, J. Zhang, J. Chung, P. Wilder-Smith, M. Brenner, J. Nelson, and Z. Chen, “Advances in oral cancer detection using optical coherence tomography,” *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 11, no. 4, pp. 811–817, 2005.
- [9] S. Kent, “Diagnosis of oral cancer using genetic programming,” Brunel University, Uxbridge, Middlesex, UB8 3PH, UK, Tech. Rep. CSTR-96-14 ; CNES-96-04, Jul. 1996. [Online]. Available: <http://citeseer.ist.psu.edu/cache/papers/cs/733/httpzSzzSzwwww.brunel.ac.ukzSz~cspgsskzSzddocumentszSztech-reportszSzCNES-96-04.pdf/diagnosis-of-oral-cancer.pdf>
- [10] M. Krishnan, P. Shah, A. Choudhary, C. Chakraborty, R. Paul, and A. Ray, “Textural characterization of histopathological images for oral sub-mucous fibrosis detection.” *Tissue Cell*, vol. 43, no. 5, pp. 318–30, 2011.
- [11] G. Hamarneh, A. Chodorowski, and T. Gustavsson, “Active contour models: application to oral lesion detection in color images,” in *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, vol. 4, 2000, pp. 2458–2463 vol.4.
- [12] Y.-n. Sun, Y.-y. Wang, S.-c. Chang, L.-w. Wu, and S.-t. Tsai, “Color-based tumor tissue segmentation for the automated estimation of oral cancer parameters,” *Microscopy Research and Technique*, vol. 73, no. 1, pp. 5–13, 2010. [Online]. Available: <http://dx.doi.org/10.1002/jemt.20746>
- [13] N. Sharma, “Comparing the performance of data mining techniques for oral cancer prediction,” in *Proceedings of the 2011 International Conference on Communication, Computing & Security*, ser. ICCCS '11. New York, NY, USA: ACM, 2011, pp. 433–438. [Online]. Available: <http://doi.acm.org/10.1145/1947940.1948029>

- [14] M. Krishnan, V. Venkatraghavan, U. Acharya, M. Pal, R. Paul, L. Min, A. Ray, J. Chatterjee, and C. Chakraborty, "Automated oral cancer identification using histopathological images: a hybrid feature extraction paradigm." *Micron*, vol. 43, no. 2-3, pp. 352–64, 2012.
- [15] A. Chodorowski, C. Choudhury, and T. Gustavsson, "Image analysis and cadx system for mucosal lesions," in *BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on*, 2008, pp. 1–4.
- [16] M. S. T. Tatyana A Zhukov, Dansheng Song, "Methods and apparatus for diagnosis and/or prognosis of cancer," Patent US 7062320, 07 12, 2012. [Online]. Available: <http://www.patents.com/us-20120177280.html/>
- [17] T. Chan and L. Vese, "Active contours without edges," *Image Processing, IEEE Transactions on*, vol. 10, no. 2, pp. 266–277, 2001.
- [18] C. Sagiv, N. Sochen, and Y. Zeevi, "Integrated active contours for texture segmentation," *Image Processing, IEEE Transactions on*, vol. 15, no. 6, pp. 1633–1646, 2006.
- [19] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Communications on Pure and Applied Mathematics*, vol. 42, no. 5, pp. 577–685, 1989. [Online]. Available: <http://dx.doi.org/10.1002/cpa.3160420503>
- [20] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations," *J. Comput. Phys.*, vol. 79, no. 1, pp. 12–49, Nov. 1988. [Online]. Available: [http://dx.doi.org/10.1016/0021-9991\(88\)90002-2](http://dx.doi.org/10.1016/0021-9991(88)90002-2)
- [21] R. Malladi, J. Sethian, and B. Vemuri, "Shape modeling with front propagation: a level set approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 2, pp. 158–175, 1995.
- [22] M. Sakalli, K.-M. Lam, and H. Yan, "A faster converging snake algorithm to locate object boundaries," *Image Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1182–1191, 2006.

- [23] M. Leventon, O. Faugeras, W. E. L. Grimson, and I. Well, W.M., "Level set based segmentation with intensity and curvature priors," in *Biomedical Imaging, 2002. 5th IEEE EMBS International Summer School on*, 2002, pp. 8 pp.-.
- [24] M. Sussman, A. S. Almgren, J. B. Bell, P. Colella, L. H. Howell, and M. L. Welcome, "An adaptive level set approach for incompressible two-phase flows," *J. Comput. Phys*, vol. 148, pp. 81–124, 1998.
- [25] C. Li, C. Xu, C. Gui, and M. D. Fox, "Level set evolution without re-initialization: A new variational formulation," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, ser. CVPR '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 430–436. [Online]. Available: <http://dx.doi.org/10.1109/CVPR.2005.213>
- [26] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.
- [27] D. Clausi and M. Jernigan, "A fast method to determine co-occurrence texture features," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 36, no. 1, pp. 298–300, 1998.
- [28] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 172–179, Jun. 1975. [Online]. Available: [http://dx.doi.org/10.1016/s0146-664x\(75\)80008-6](http://dx.doi.org/10.1016/s0146-664x(75)80008-6)
- [29] A. Chu, C. M. Sehgal, and J. F. Greenleaf, "Use of gray value distribution of run lengths for texture analysis," *Pattern Recogn. Lett.*, vol. 11, no. 6, pp. 415–420, Jun. 1990. [Online]. Available: [http://dx.doi.org/10.1016/0167-8655\(90\)90112-F](http://dx.doi.org/10.1016/0167-8655(90)90112-F)
- [30] B. V. Dasarathy and E. B. Holder, "Image characterizations based on joint gray level-run length distributions," *Pattern Recogn. Lett.*, vol. 12, no. 8, pp. 497–502, Aug. 1991. [Online]. Available: [http://dx.doi.org/10.1016/0167-8655\(91\)80014-2](http://dx.doi.org/10.1016/0167-8655(91)80014-2)

- [31] X. Sun, S.-H. Chuang, J. Li, and F. McKenzie, "Automatic diagnosis for prostate cancer using run-length matrix method," pp. 72 603H–72 603H–8, 2009. [Online]. Available: [+http://dx.doi.org/10.1117/12.811414](http://dx.doi.org/10.1117/12.811414)
- [32] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: <http://dl.acm.org/citation.cfm?id=944919.944968>
- [33] G. Forman, I. Guyon, and A. Elisseeff, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.
- [34] J. Fan, R. Samworth, and Y. Wu, "Ultrahigh dimensional feature selection: Beyond the linear model," *J. Mach. Learn. Res.*, vol. 10, pp. 2013–2038, Dec. 2009. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1577069.1755853>
- [35] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.
- [36] E. R. Tufte, *The visual display of quantitative information*. Cheshire, CT, USA: Graphics Press, 1986.
- [37] C. Kamath, *Scientific Data Mining: A Practical Perspective*. 3600 Market Street, 6th Floor, Philadelphia, PA: SIAM, 2009.
- [38] B. Caby, S. Kieffer, M. de Saint Hubert, G. Cremer, and B. Macq, "Feature extraction and selection for objective gait analysis and fall risk assessment by accelerometry," *Biomedical engineering online*, vol. 10, no. 1, pp. 1–1, 2011. [Online]. Available: <http://dx.doi.org/10.1186/1475-925X-10-1>
- [39] I. K. Fodor and C. Kamath, "Dimension reduction techniques and the classification of bent double galaxies," *Comput. Stat. Data Anal.*, vol. 41, no. 1, pp. 91–122, Nov. 2002. [Online]. Available: [http://dx.doi.org/10.1016/S0167-9473\(02\)00061-0](http://dx.doi.org/10.1016/S0167-9473(02)00061-0)
- [40] M. Vasconcelos and N. Vasconcelos, "Natural image statistics and low-complexity feature selection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 228–244, 2009.

- [41] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, Tech. Rep., 2010.
- [42] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Dec. 2004. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1005332.1044700>
- [43] N. Garcia-Pedrajas and D. Ortiz-Boyer, "Improving multiclass pattern recognition by the combination of two strategies," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 6, pp. 1001–1006, 2006.
- [44] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, 1995, pp. 273–297.
- [45] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [46] J. E. Jackson, *A User's Guide to Principal Components*, ser. Wiley Series in Probability and Statistics. Wiley-Interscience, Sep. 2003. [Online]. Available: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471471348>
- [47] J. Yang and J. Yu Yang, "Why can LDA be performed in PCA transformed space?" *Pattern Recognition*, vol. 36, no. 2, pp. 563 – 566, 2003, doi:10.1016/S0031-3203(02)00048-1. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0031320302000481>
- [48] L. Wang, C. Shen, and R. Hartley, "On the optimality of sequential forward feature selection using class separability measure," in *Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on*, 2011, pp. 203–208.
- [49] Y. Liu and Y. Zheng, "Fs sfs: a novel feature selection method for support vector machines," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 5, 2004, pp. V-797–800 vol.5.

- [50] D. Wang, H. Zhang, R. Liu, and W. Lv, "Feature selection based on term frequency and t-test for text categorization," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, ser. CIKM '12. New York, NY, USA: ACM, 2012, pp. 1482–1486. [Online]. Available: <http://doi.acm.org/10.1145/2396761.2398457>
- [51] I. Kononenko, E. Simec, and M. Robnik-Sikonja, "Overcoming the myopia of inductive learning algorithms with relieff," *Applied Intelligence*, vol. 7, pp. 39–55, 1997.
- [52] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Mach. Learn.*, vol. 53, no. 1-2, pp. 23–69, Oct. 2003. [Online]. Available: <http://dx.doi.org/10.1023/A:1025667309714>
- [53] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *Proceedings of the European conference on machine learning on Machine Learning*, ser. ECML-94. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1994, pp. 171–182. [Online]. Available: <http://dl.acm.org/citation.cfm?id=188408.188427>
- [54] C. M. Florkowski, "Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests." *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists*, vol. 29 Suppl 1, Aug. 2008. [Online]. Available: <http://view.ncbi.nlm.nih.gov/pubmed/18852864>
- [55] D. G. Altman and J. M. Bland, "Statistics notes: Diagnostic tests 1: sensitivity and specificity," *BMJ*, vol. 308, no. 6943, p. 1552, 6 1994.
- [56] G. Ou, Y. Murphey, and L. Feldkamp, "Multiclass pattern classification using neural networks," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 4, 2004, pp. 585–588 Vol.4.
- [57] S. Tso, X. Gu, and W. Zhang, "An index-based classification scheme using neural networks for multiclass problems," in *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, vol. 3, 1998, pp. 1899–1904 vol.3.

Image Database

1. Carcinoma buccal mucous lower GB sulcus : Ulceroproliferative growth
2. Carcinoma Palate : Ulceroproliferative growth
3. Carcinoma Tongue : lateral border of tongue , ulcerative growth
4. Leucoplakia : white patch over buccal mucosa , a premalignant condition
5. Carcinoma palate : ulceroproliferative growth
6. Erythroplasia : Red patch , a premalignant condition
7. Carcinoma Buccal mucosa : Ulceroproliferative growth
8. Carcinoma Buccal mucosa : Ulceroproliferative growth
9. Carcinoma Retromolar area
10. Carcinoma retromolar area
11. Leucoplakia over lower lip
12. Verrucous carcinoma buccal mucosa
13. Verrucous carcinoma buccal mucosa
14. Verrucous carcinoma buccal mucosa
15. Carcinoma Tongue : lateral border
16. Carcinoma lower lip : ulcerative growth
17. Carcinoma tongue : ulcerative growth

18. Carcinoma angle of mouth : ulcero-proliferative growth
19. Carcinoma angle of mouth : ulcero-proliferative growth
20. Case of Melanoplakia and Leukoplakia

Database contains 51 Normal cases also along with the above cases



Formulae Used

TABLE 1: First Order Statistical Features

| Feature Name | Formula |
|--------------------|--|
| Minimum | $I_{min} = \min \{I(x, y)\}$ |
| Maximum | $I_{max} = \max \{I(x, y)\}$ |
| Mean | $\mu = \frac{1}{XY} \sum_{x=1}^X \sum_{y=1}^Y I(x, y)$ |
| Variance | $\sigma^2 = \frac{1}{XY-1} \sum_{x=1}^X \sum_{y=1}^Y [I(x, y) - \mu]^2$ |
| Standard Deviation | $\sigma = \sqrt{\text{Variance}}$ |
| Skewness | $\frac{1}{XY} \sum_{x=1}^X \sum_{y=1}^Y \left[\frac{I(x, y) - \mu}{\sigma} \right]^3$ |
| Kurtosis | $\frac{1}{XY} \sum_{x=1}^X \sum_{y=1}^Y \left[\frac{I(x, y) - \mu}{\sigma} \right]^4$ |

TABLE 2: GLCM Features

| Feature Name | Formula |
|-------------------|--|
| Contrast | $\sum_{i=1}^N \sum_{j=1}^N i - j ^2 p(i, j)$ |
| Correlation | $\sum_{i=1}^N \sum_{j=1}^N \frac{(i - \mu_i)(j - \mu_j) p(i, j)}{\sigma_i \sigma_j}$ |
| ASM(Energy) | $\sum_{i=1}^N \sum_{j=1}^N \{p(i, j)\}^2$ |
| Homogeneity | $\sum_{i=1}^N \sum_{j=1}^N \frac{p(i, j)}{1 + i - j }$ |
| Entropy | $\sum_{i=1}^N \sum_{j=1}^N [(p(i, j) \log(p(i, j)))]$ |
| SoS-Variance | $\sum_{i=1}^N \sum_{j=1}^N [(i - \mu)^2 p(i, j)]$ |
| IDM | $\sum_{i=1}^N \sum_{j=1}^N \left(\frac{1}{1 + (i - j)^2} p(i, j) \right)$ |
| Sum Average(SA) | $\sum_{i=2}^{2N} [ip_{x+y}(i)]$ |
| Sum Variance(SV) | $\sum_{i=2}^{2N} [(i - SA)^2 p_{x+y}(i)]$ |
| Sum Entropy(SE) | $\sum_{i=2}^{2N} p_{x+y}(i) \log [p_{x+y}(i)]$ |
| Diff. Entropy(DE) | $\sum_{i=0}^{N-1} p_{x-y}(i) \log [p_{x-y}(i)]$ |

Where

$p(i,j)$ = The Normalized Gray Level Co-occurrence Matrix

N = The number of discrete gray levels of image

$$\mu_x = \sum_{i=1}^N \left(i * \sum_{j=1}^N p(i,j) \right)$$

$$\mu_y = \sum_{j=1}^N \left(j * \sum_{i=1}^N p(i,j) \right)$$

$$\sigma_x = \sum_{i=1}^N \left((i - \mu_x)^2 * \sum_{j=1}^N p(i,j) \right)$$

$$\sigma_y = \sum_{j=1}^N \left((j - \mu_y)^2 * \sum_{i=1}^N p(i,j) \right)$$

μ_x, μ_y = mean of p_x and p_y

σ_x, σ_y = Std. Deviation of p_x and p_y

where $p_x(i)$ and $p_y(j)$ are row and column marginal probabilities

$$p_x(i) = \sum_{j=1}^N p(i,j)$$

$$p_y(j) = \sum_{i=1}^N p(i,j)$$

$$p_{x+y}(k) = \sum_{i=1}^N \sum_{j=1}^N p(i,j) \quad , \quad \text{where } k = 2, 3, \dots, 2N \quad \text{and } i + j = k$$

$$p_{x-y}(k) = \sum_{i=1}^N \sum_{j=1}^N p(i,j) \quad (1)$$

Where $k=0,1,2,\dots,N-1$ and $|i - j| = k$

TABLE 3: GLRL Features

| Feature Name | Formula |
|------------------------------------|---|
| Short Run Emphasis | $SRE = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N \left[\frac{P(i,j)}{j^2} \right]$ |
| Long Run Emphasis | $LRE = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N [P(i,j)] * j^2$ |
| Low Gray-level Run Emphasis | $LGRE = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N \left[\frac{P(i,j)}{i^2} \right]$ |
| High Gray-level Run Emphasis | $HGRE = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N [P(i,j)] * i^2$ |
| Short Run Low Gray-level Emphasis | $SRLGE = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N \left[\frac{P(i,j)}{i^2 * j^2} \right]$ |
| Short Run High Gray-level Emphasis | $SRHGE = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N \left[\frac{P(i,j) * i^2}{j^2} \right]$ |
| Long Run Low Gray-level Emphasis | $LRLGE = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N \left[\frac{P(i,j) * j^2}{i^2} \right]$ |
| Long Run High Gray-level Emphasis | $LRHGE = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N [P(i,j)] * i^2 * j^2$ |
| Gray-level Non-Uniformity | $GLN = \frac{1}{n} \sum_{i=1}^M \left(\sum_{j=1}^N P(i,j) \right)^2$ |
| Run Length Non-Uniformity | $RLN = \frac{1}{n} \sum_{j=1}^N \left(\sum_{i=1}^M P(i,j) \right)^2$ |
| Run Percentage | $RP = \frac{n}{M * N}$ |

where

$p(i,j)$ = total number of runs with pixel gray level of i and run length of j

n = no. of pixels in image.

M = highest gray level and N = max. run length possible.