Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

COMPUTER AIDED ANALYSIS OF SKIN LESIONS

by
Sukanya Panja

Effective screening to detect the skin cancer accurately in the early stage is essential for reducing the mortality of skin cancer. Surface features, such as texture and pigmentation area from the surface, epi- illumination images of the skin lesions have been well correlated to detect skin cancer. An increase in the lesion's subsurface blood volume has been correlated to early diagnosis of malignant melanoma. A method for estimating the optimal features is obtained. The optimal features help in accurately classify the skin lesion in various grades. To make the process faster these optimal features are clustered. The optimal clusters are obtained by genetic algorithm. The optimal cluster centers act as input to the SVM classifier and the kernel parameters are obtained. Finally, parameters of the kernel function are optimized by genetic algorithm, which help in classifying the skin lesions into various grades leading to early diagnosis of skin cancer.

COMPUTER AIDED ANALYSIS OF SKIN LESIONS

by
Sukanya Panja

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree
of Master of Science in Electrical Engineering

Department of Electrical and Computer Engineering

January 2015

APPROVAL PAGE

COMPUTER AIDED ANALYSIS OF SKIN LESIONS

Sukanya Panja

Dr. Atam. P. Dhawan, Thesis Advisor — Date
Distinguished Professor of Electrical and Computer Engineering, NJIT

Dr. Yun-Qing Shi, Committee Member — Date
Professor of Electrical and Computer Engineering, NJIT

Dr. Sui-Hoi Edwin Hou, Committee Member — Date
Assistant Professor of Electrical and Computer Engineering, NJIT

BIOGRAPHICAL SKETCH

Author: Sukanya Panja

Degree: Master of Science

Date: January 2015

Undergraduate and Graduate Education:

- Master of Science in Electrical Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2015

- Bachelor of Technology in Electrical and Electronics Engineering,
  Sikkim Manipal Institute of Technology, Sikkim, India, 2011

Major: Intelligent Systems

.

To my brother, Sayan Banerjee who has been my biggest inspiration

ACKNOWLEDGMENT

My deepest thanks to my advisor, Dr. Atam P. Dhawan for all his guidance, support, and encouragement throughout my thesis work. I have learned a tremendous amount under his supervision, for which I am deeply grateful. His knowledge and assistance has been invaluable. I would also like to thank my committee members: Dr. Yun-Qing Shi and Dr. Sui-Hoi Edwin for their feedback.

I would also like to express my gratitude to Dr. Brian D'Alessandro who have assisted and guided me in the initial stage of the thesis. I am also grateful to my brother, Aminur Rahman who have constantly supported me and helped me to understand the concepts more clearly.

Lastly, I could not have done this without the support of my family and friends especially Dipika Ghosh Dostidar, Sharon Abraham John, Gyan Ranjan Sahu and Rounaq Gandhi.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

Table **Page**

LIST OF FIGURES

LIST OF SYMBOLS

$\int$	Integration
∂	Partial Differential

LIST OF DEFINITIONS

Accuracy	How closely an instrument measures the true or a of the process variable being measured or sensed.
Optimal	Best or most favorable.
Classification	The action or process of classifying something according to shared qualities or characteristics

CHAPTER 1

INTRODUCTION

1.1 Motivation

Skin cancer is the most common form of cancer. More than 3.5 million cancer cases are diagnosed annually. Most type of skin cancer are curable but the deadliest of them, melanoma is expected to result over 8700 deaths. Even though melanoma only represents 3% of the skin cancer cases, it results over 75% of the skin cancer deaths. Thus early detection and diagnosis of melanoma is crucial to treating malignancy and preventing deaths. Likewise, an efficient method of screening is essential, as one individual may have many lesions to be analyzed. Such analysis must be accurate in a time efficient manner.

Melanocytic nevi, commonly known as moles are skin lesions which develop from the melanocytes in the skin. Skin melanocytes produce melanin which is mainly dark in color. While there are different types of moles, two most common types are junctional nevus and compound nevus. In junctional nexus, melanocytes occur in the epidermal layer of the skin and are found all the way down into the dermis as well. The vast majority of the nevi are benign but in some individual the nevi grows malignant over time. Lesions must also necessarily grow into the dermal layer in order to be diagnosed as malignant. Therefore, it is often difficult to distinguish a compound nexus from a malignant melanoma because both penetrate the dermis. Once in the dermis, the melanoma first metastasizes, first to the lymphatic system and blood stream, then finally to the rest of the body.

The problem of early detection of melanoma is difficult for normal people as well for the physicians. The accuracy rate of diagnosis falls further for the non- experts who do not specialize in early melanoma detection. Conventional analysis for the detection of the melanoma is done by the conventional 'ABCD' rule where 'ABCD' denotes:

(A)symmetry , (B)order irregularity, (C)olor varigation and (D)iameter generally >6 mm. Instruments such as Dermite have been used for with surface lighting and magnification to analyze the visible structure of a nevus. Deeper subsurface information, such as subcutaneous pigmentation, depth of invasion and indications of increased blood flow are critical factors in early melanoma detection. As a result, much of the effort is being put into the evaluation of novel noninvasive optical imaging techniques as a way to detect and analyze the morphological changes thereby improving the patient diagnosis accuracy with minimal need for invasive and time consuming biopsy procedures.

1.2 Overview

To this end, an interactive segmentation tool is presented which uses a contribution of k-means clustering, wavelet analysis and morphological operations to segment the lesion pigmentation and blood volume area visible to 2D TLM and ELM images. The user is then presented with six segmentation suggestion for both images. The ratio of the TLM segmented area to ELM segmented area is investigated as an indicator of dysplasia in skin lesion for early detection of skin cancer. In addition to the ratio, a set of texture and color features are extracted from the ELM and TLM images based on wavelet analysis, second order histogram analysis and boundary characteristics. Using this set of features,

classification is performed to classify the lesions into three groups: mild, moderate/severe and malignant melanoma.

There are a lot of features which might affect the analysis of the skin lesion. The optimal feature set is obtained by passing the features through a search process called the genetic algorithm. The optimal feature set helps in increasing the accuracy of the classification of the lesion. To further increase the accuracy of the classification the optimal features are clustered. The optimal clusters are obtained through genetic algorithm. Clustering can also be done through K-means but k-means has a drawback that it converges to local minima. This problem is overcome by genetic algorithm as in genetic algorithm there is a minimal chance of converging to a local minima. The optimal clusters contain homogeneous data which can be classified more accurately.

Next, the error of misclassification is minimized by optimizing the kernel parameters and feature subset. This is done by the genetic algorithm simultaneously. This further increases the accuracy of the classification.

1.2 Objectives

The overall objectives of this work are:

1. To develop and implement two dimensional texture features based on white light ELM and TLM images with an appropriate selection algorithm for classifying the skin lesions into various grades of severity

2. To obtain the optimal clusters which will optimally cluster the features which are of homogeneous nature.

3. To optimize the parameters of the kernel function and also the feature subset simultaneously this will help in increasing the accuracy of the classifier for analyzing the skin lesion into various grades of severity.

The goal of this work is to increase the accuracy of the classifier which will accurately classify the skin lesion into various grades of severity and also make the process faster. This classification is done using real, clinical, pathologically validated lesions as opposed to relying on simulations alone.

CHAPTER 2

IMAGE PROCESSING

One of the methods which are used for classification of the skin lesion involves color and texture analysis of the epi- illuminated and trans- illuminated images. In malignant lesions, the vascularity of the skin increases. In this chapter, an interactive segmentation is applied on the epi- illumination images and trans-illuminated images. In skin lesion, the ratio of the segmented TLM area to the segmented ELM area gives the indication of vascularity. Various color and texture features are extracted from the image. These features along with the TLM to ELM ratio are used in a genetic algorithm to get the optimal number of clusters. Once the optimal numbers of clusters are obtained, they are then fed to the learning based classifier to classify whether the skin lesion is mild, moderate severe or of malignant categories. The optimal number of features and the clusters would help in improving the results.

2.1 Artifact Removal

Like many of the real world images, the images that are acquired are not perfect. Most of the images contain a lot of black marks. These black marks are mostly made by pen or marker by the dermatologist indicating the area to be imaged. These black marks can be mistaken to be lesion by the segmentation algorithm and thus needs to be detected and removed first before any subsequent lesion analysis.

The black marks are detected by wavelet analysis. . Wavelet analysis is a powerful tool which decomposes an image into variable spatial and frequency resolutions. At a single

level of decomposition the signal is filtered into high and low components which provided with a set of approximation coefficients from the low pass filter and a set of detail components from the high pass filter. In this analysis, Debauchees D4 wavelet is being used for the decomposition in both the spatial dimensions of the digital image. For the ELM image, the artifacts are removed by using the thresholding and the morphological processing of the high-high wavelet coefficients of the saturation image. The ELM image is first converted from the red, green, blue (RGB) color space to the hue, saturation value (HSV) color space. Then the thresholding and morphological processing is applied to segment the artifact from the image. Likewise, the artifacts in the TLM image is segmented through similar thresholding and morphological processing of the high – high wavelet coefficients of the R- channel in the TLM image.

Figure 2.1 Artifact Segmented in Lesion image.

2.2 Background and Color Correction of TLM Image

The original TLM image is largely saturated with red color which makes it difficult to distinguish the features and determine the blood intensity in the image. In order to overcome this problem the image need to be corrected. This correction is done by segmenting the image into two classes: a background class and a foreground class. The pixels of the background are fit into a two dimensional, second – order polynomial curve. The curve represents the background profile without the lesion. The estimation of the background without the lesion helps in correction of the TLM image with the lesion. The background estimate helps in correcting the TLM image with respect to the background by flattening out the intensity of the skin surrounding the lesion and heightening the contrast between the lesion and the surrounding skin.

$$\hat{I}_{\mathrm{R}} = \left(\frac{I_{\mathrm{R}}}{BG_{\mathrm{R}}} - 0.16 \right)^{1.6} \tag{2.1}$$

$$\hat{I}_{\mathrm{G}} = \left(\frac{I_{\mathrm{G}}}{3 \cdot BG_{\mathrm{G}} + 0.02} \right)^{0.452} \tag{2.2}$$

Where I_R and I_G are the original red and green channels of the TLM image, respectively and BG_G, BG_R are the estimated red and green channel background curves.

The exponents used in the above equation helps in color and contrast correction. Thus the final corrected TLM image is more easily segmented and is more visually informative to the user.

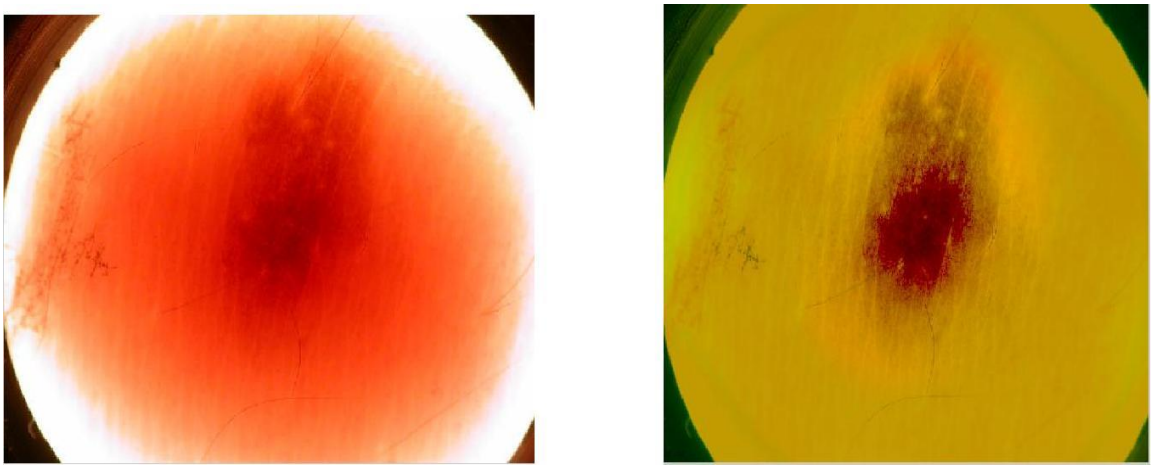

Figure 2.2 Left**:** Original TLM image. Right**:** Background and color corrected TLM image

CHAPTER 3

EXTRACTION OF FEATURES

The features for the skin lesion are obtained from the ELM and TLM images. Three out of the six segmentation options for ELM image and five of the six TLM images obtained were based on k-means clustering analysis. K-means clustering is a clustering technique. In this technique cluster centroids are randomly chosen from the data set. Each data element is then assigned to a cluster. The assignment of data elements to the cluster is done by obtaining the Euclidean distance between the data and the cluster center. The shortest distance obtained between the data and the cluster centroid, is the cluster to be chosen. The k-means algorithm then recalculates the centroid location of each cluster and updates the cluster assignment of the data based on the new cluster centroids. In the ELM image, the algorithm was run on the saturation component of the HSV representation of ELM image and for the TLM image, the algorithm runs on the G channel of the RGB image.

Wavelet analysis was used to produce the additional features. Three additional segmentation options were produced by wavelet analysis for ELM image and one additional image for TLM images. To obtain the additional features via wavelet analysis, from ELM image k-means algorithm was run on the high-high decomposition of the saturation channel to produce three levels of segmentation options. In the TLM image, one additional feature is obtained by k-means algorithm. The k- means algorithm is run on the high-high

decomposition of the G-channel to produce one feature for the TLM image.

Once the TLM and the ELM images are segmented with respect to their pigmentation area and blood surface area, the ratio between the areas of the TLM to the ELM is calculated. This area gives the indication of the vascularity. Increase in the vascularity is an indication of the malignant lesion.

3.1 ADDITIONAL FEATURES

A number of additional features are obtained from the image which helps in classification of the lesion based on texture and morphological characteristics.

3.1.1 Wavelet Features

Wavelet decomposition using D4 wavelet was used to obtain the wavelet features of the image. One level of decomposition results in a set of detail (H) and approximation (L) coefficients for each of the two dimension of the image. Thus, four sub bands are generated: LL, LH, HL and HH. Each successive level of the decomposition further decomposes each of the sub bands into sub-sub bands. Hence, we can say that for any level, 4 sub-bands are created. On each sub-band two features are computed.

1. Energy:

$$\frac{1}{N}\sum_{u,v} c_{u,v}^2 \qquad \textbf{(3.1)}$$

2. Entropy:

$$-\sum_{u,v}^{N}\left[\frac{c_{u,v}^{2}}{\sum_{u,v}c_{u,v}^{2}}\log_{10}\left(\frac{c_{u,v}^{2}}{\sum_{u,v}c_{u,v}^{2}}\right)\right] \quad (3.2)$$

These features are obtained from first, second and third level of decomposition of R and G channels of TLM images and S and V channels of ELM image .A total of $2.4.4^{1}+2.4.4^{2}+2.4.4^{3}=672$ wavelet features.

Since there are too many features for efficient feature selection and undoubtedly it contains many correlated features hence, independent component analysis (ICA) was used to reduce the wavelet feature set. ICA seeks to find out a linear representation of the given data such that the feature components are statistically as independent as possible. In ICA, feature components with small eigenvalues can be discarded to reduce the dimensionality of the feature set. For a set of 160 wavelet features, the dimension is reduced to 16 features using ICA.

3.1.2 GLCM Features

The second type of features which are being used for analysis of the image are obtained from the texture analysis. Texture analysis utilizes second order histogram information which is also known as grey level co-occurrence matrix (GLCM).GLCM features are obtained from the grey level difference between two consecutive pixels separated by a given distance. The image is first scaled down to eight grey levels, so that the size of the GLCM is 8x8, where the rows indicate the value of first pixel and the column correspond to the second pixel. For each pair of pixel in the image, the entry in the GLCM is

incremented. Thus, the GLCM acts like a counter which counts the number of pixels with given intensity values and separated by a distance d.

Seven features are obtained from the GLCM, which characterizes the texture of the original image:

1. Contrast:

$$\sum_{i,j} |i-j|^2 p_{ij} \tag{3.3}$$

2. Correlation:

$$\sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)p_{ij}}{\sigma_i \sigma_j} \tag{3.4}$$

where μ_i, μ_j, σ_i, σ_j are the mean and the standard deviation of p_i and p_j

3. Energy:

$$\sum_{i,j} p_{ij}^2$$

4. Homogeneity:

$$\sum_{i,j} \frac{p_{ij}}{1+|i-j|} \tag{3.5}$$

5. Angular-second moment:

$$\sum_{i,j} p_{ij}^2 \tag{3.6}$$

6. Entropy:

$$-\Sigma_i \Sigma_j p_{ij} \cdot \log \cdot p_{ij} \quad (3.7)$$

7. Mean:

$$\sum_{i=j=0}^{N-1} ij(p_{ij}) \quad (3.8)$$

These features are obtained from the R and G channels of the TLM image in RGB representation and S and V channels of the ELM image in the HSV representation. To obtain the features an offset value of d = 1, 2, 3, 4, 5 were used in four different directions (up, down, left, right). Each of the direction gives different GLCM values, but the computed GLCM matrix consists of the average of the values across four direction. Hence, a total of 4*7*5=140 GLCM features are generated.

The GLCM feature set generated is too large. It is very well clear that all the features cannot be used for significant analysis of the skin lesion. Thus, the most significant features are obtained through the Independent component analysis (ICA). ICA is used to reduce the number of feature down to 28 most independent features.

3.1.3 Region Boundary Features

Five additional features were found based on the TLM and ELM lesion boundaries selected from the segmentation interface. For computing the region boundary features, the image is converted into a binary image. These include:

1. Eccentricity: The eccentricity gives a scalar value which indicated the eccentricity of the ellipse that has the same second moment as the boundary.

2. Solidity: It is a scalar value which specifies the proportion of the pixels in the convex hull that are also in the region. This is given by the ratio of the area of the region to the area of the region within the convex boundary.
3. Extent: It is the scalar value that specifies the ratio of the pixels in the region to the pixels in the total bounding box.
4. Perimeter: It is a scalar value that gives the distance around the boundary of the region. The perimeter is computed by calculating the distance between each adjoining pair of pixels around the region
5. Circularity: Perimeter of the region divided by the diameter of the circle whose area is equal to the area of the region

Thus, a total of 2x5 =10 region boundary features are obtained from the ELM and TLM lesion boundaries.

3.2 Normalization of Features

The features that are obtained have different range and so it is important to normalize the features so that they have the same range of values. The normalization of data helps in identifying whether the lesion is mild, severe or malignant. There are a lot of ways to normalize data, including linear scaling to unit range, linear scaling to unit variance, and transformation to uniform random variable, rank normalization, or normalization after fitting it into a given distribution. For this application, features are normalized by the linear scaling to unit variance method.

The feature normalization equation is as follows:

$$\hat{\mathrm{x}} = \frac{\mathrm{x} - \mu_{\mathbf{x}}}{\sigma_{\mathbf{x}}} \tag{3.8}$$

Where μ_x and σ_x are the mean and standard deviation of the feature vector x to be normalized. $\hat{x}$ is the resulting normalized feature vector. This normalization is applied to all the features which are obtained. The normalized features consist of TLM/ELM ratio, 16 ICA wavelet features, 28 ICA GLCM features and 10 boundary features. Thus, a total of 54 normalized ICA features are obtained.

CHAPTER 4

OPTIMAL FEATURES SELECTION

Too many features can cause the problem of over classification. Ability of good classification by good features can be clouded by noisy features which lead to poor classification of the lesion. Different features lead to different percentage of classification accuracy. Genetic Algorithm was used to obtain an optimal subset of features.

4.1 Introduction to Genetic Algorithm

Genetic Algorithm is a technique used to obtain the optimal solution of a problem which is difficult or impossible to obtain otherwise. Genetic algorithm belongs to the larger class of evolutionary algorithm which generates solution to optimization problems using techniques inspired by the natural evolution.

1. A genetic representation of the solution domain
2. A fitness function to evaluate the solution domain

Basic flowchart of genetic algorithm:

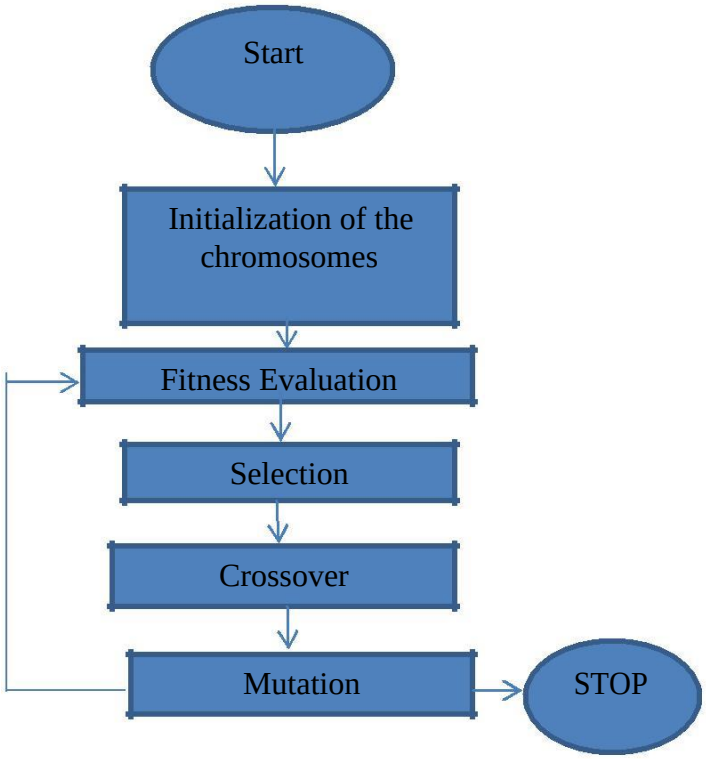

Figure 4.1 Flowchart of basic genetic algorithm.

4.1.1 Initialization

Initially, all the solutions are randomly generated to form the initial population. The population size depends on the nature of the problem and contains as many solutions as possible. Generally, the solutions are generated randomly allowing the entire range of possible solutions. Individuals in the population represent the different solutions for the problem. Information of these individuals are encoded in the chromosome.

4.1.2 Evaluation of the Fitness

The fitness value is obtained from the problem which is being dealt with. It describes the ability to reproduce and survive. Fitness is a probability rather than an actual number of

offspring. The fitness is the probability which helps in identifying whether the individual will be included among the group selected as parents or the next generation. The fitness of an individual is modeled through the use of fitness function. The fitness function takes a single individual or a single chromosome as an input and evaluates how well the problem is solved with the help of the input chromosome. The fitness function returns the fitness evaluated. Chromosomes with higher fitness have a greater chance to reproduce and pass their information to the offspring. With each generation, the population of the chromosome should converge closer to the optimal solution of the problem.

A chromosome is most often represented as a one dimensional binary string. The chromosome can be either binary or a set of real valued numbers. The actual shape of the chromosome may vary as well. Regardless of the actual format of the chromosome, the only requirements are that it contains finite set of numbers and it can evaluate the fitness function.

4.1.3 Selection

During each generation, a proportion of the existing group is selected to reproduce the next generation. The proportion of the existing group which is selected to reproduce the next generation is done with the help of the selection process. Individual solutions are selected based on the fitness value. Certain selection processes rank the fitness of each solution and preferentially select the best two. One of the selection processes is the roulette wheel selection. In this process, each chromosome has a small chance of being selected but the chromosome with higher fitness value has a higher chance of being selected.

A simple way of finding the selection probability p_j for each chromosome in the population size N with fitness evaluation [F_1,F_2,.....F_n] is:

$$p_j = \frac{F_j}{\sum_{i=1}^{N} F_i} \tag{4.1}$$

However, this technique does not perform well in populations where the fitness evaluations have a small percentage deviation from each other. Alternatively, a fitness technique known as linear normalization scales the fitness values to a set range of [0,100] before computing the selection probability. Another method ranks the fitness values in the ascending order. The selection probability of the chromosome is:

$$p_j = \frac{\text{rank}(F_j)}{0.5 \cdot N(N+1)} \tag{4.2}$$

This method sorts the probability in ascending order with the least fit to most fit chromosome.

4.2 Genetic Operators

The next step is to generate the second generation of population from those selected through a combination of genetic operators: crossover and mutation. Each new solution to be produced, a pair of parent solutions is selected for reproduction from the pool selected previously. A child solution is obtained from the crossover and mutation which is created from the parent solutions. New parents are selected for each new child and the process continues until a new population of solutions of appropriate size is not generated. These processes ultimately generate the next generation of population which is different from the initial population and the average fitness will be increased for the next

generation. There are a number of reproduction techniques which includes generational replacement and steady-state reproduction. In generational replacement, the entire population is replaced by the offspring in each generation while in steady –state reproduction one or two chromosomes are replaced by their offspring in each generation. Therefore, it is clear that steady –state reproduction requires larger time to converge than generational replacement. However, when we compare the performance of the two methods, steady-state reproduction produces better results than generational reproduction.

4.2.1 Crossover

Once the selection probability of each chromosome is obtained, two parents are picked based on the two best probabilities. These two best chromosomes are selected for reproduction. Reproduction is done by “crossover”. Crossover is a method by which two parent chromosomes are mixed to produce two new offspring in their place. The simplest method of cross over is one point crossover.

Each parent that is chromosome A and Chromosome B are divided into two sections.

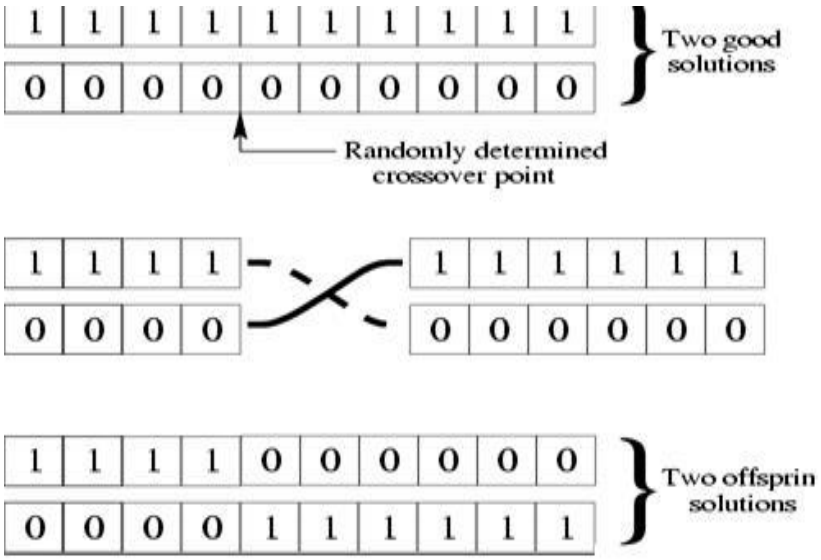

Figure 4.2 Single point crossover.

4.2.2 Mutation

Another genetic operator is Mutation. Within each generational loop, each chromosome in the population has the probability of randomly changing one of its values. In binary chromosomes, one bit is randomly flipped while in real valued chromosomes, a new random number can be replaced on one of the positions in the chromosome. Mutation adds a new generic element into the chromosome. This, helps in converging to the global minima as mutation allows the population to explore small changes around the solution to find the optimal solution within the given solution neighborhood.

4.3 Termination Condition

Common termination conditions are:

1. A solution is found that satisfies the minimum criteria
2. Fixed number of generation reached
3. Allocated budget reached

4.4 Genetic Algorithm in Feature Selection

In feature selection, each chromosome in the GA is defined to be a binary array of length equal to the number of features. In this way, each bit in the chromosome corresponds to individual features. Each bit in the chromosome represent if that feature is present in the final classification or not and hence, if the bit value in the chromosome is 1, it means that the feature is present while if the bit value is 0 that means that the feature is absent in the chromosome.

The fitness function contains a training algorithm which computes how well the classification is performed by the chromosome. Reproduction techniques include steady-state reproduction. The fitness function provides the accuracy of the SVM classifier using the features from the chromosome. This accuracy may vary over multiple trials of the evaluation. Nevertheless, the SVM classifier was run 50 times and the average accuracy was taken as the chromosome fitness value.

For the parent selection technique, a roulette wheel selection was used with rank normalization on the fitness. Uniform crossover and random binary mutation with a

crossover rate of 70% and mutation rate of 20%. The GA was run with steady-state replacement for 100 generations to find the optimal subset of features.

4.5 Optimal Feature Selection Using Genetic Algorithm

Initially, the total numbers of features selected were 38 and the accuracy of the training set was 92% and the accuracy for the test 81%. The fitness value was 77. Now, the optimal number of features selected so that the training and test accuracy increases along with the fitness value. In this, a lot of features were added to test the accuracy of the training and the test set along with the fitness value.

A lot of trials were conducted and in the end the optimal set of 56 features are obtained.

The optimal feature set is as follows:

1. TLM/ELM Area ratio
2. ELM Area
3. Wavelet ICA 1
4. Wavelet ICA 2
5. Wavelet ICA 3
6. Wavelet ICA 4
7. Wavelet ICA 5
8. Wavelet ICA 6
9. Wavelet ICA 7
10. Wavelet ICA 8
11. Wavelet ICA 9
12. Wavelet ICA 10
13. Wavelet ICA 11
14. Wavelet ICA 12
15. Wavelet ICA 13
16. Wavelet ICA 14
17. Wavelet ICA 15
18. Wavelet ICA 16
19. GLCM ICA 1
20. GLCM ICA 2
21. GLCM ICA 3
22. GLCM ICA 4
23. GLCM ICA 5
24. GLCM ICA 6
25. GLCM ICA 7
26. GLCM ICA 8
27. GLCM ICA 9
28. GLCM ICA 10
29. GLCM ICA 11
30. GLCM ICA 12
31. GLCM ICA 13
32. GLCM ICA 14
33. GLCM ICA 15
34. GLCM ICA 16
35. GLCM ICA 17
36. GLCM ICA 18
37. GLCM ICA 19
38. GLCM ICA 20
39. GLCM ICA 21
40. GLCM ICA 22
41. GLCM ICA 23
42. GLCM ICA 24
43. GLCM ICA 25
44. GLCM ICA 26
45. GLCM ICA 27
46. GLCM ICA 28
47. TLM Eccentricity
48. TLM Solidity
49. TLM Extent
50. TLM Perimeter
51. TLM Circularity
52. ELM Eccentricity
53. ELM Solidity
54. ELM Extent
55. ELM Perimeter
56. ELM Circularity

The 56 features give the best optimization curve. The optimization curve is given by:

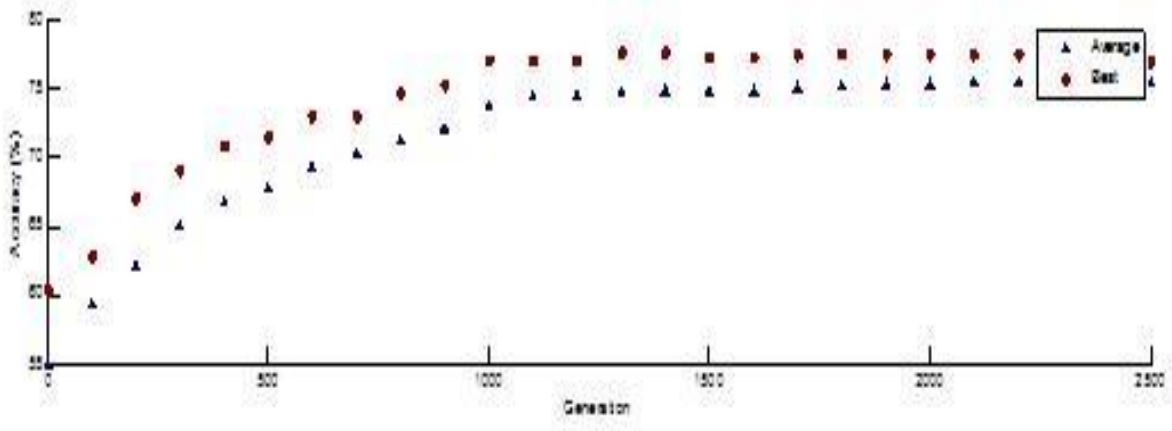

Figure 4.3 Optimization curve for selected best features.

These optimal features give a training accuracy of 96%, test accuracy of 84.6% and the fitness value of 78.03 is obtained. This result may vary with every run.

CHAPTER 5

Relation of Specificity and Sensitivity with Parameters

5.1 Specificity

Specificity is a statistical measure used to calculate the performance of a binary classification test. Specificity, which is also called true negative rate, measures the total number of negative cases which are correctly identified. For example, percentage of healthy people being identified as not having the disease. A perfect predictor would have 100% specificity such that all healthy people are being predicted as healthy.

For example, a medical test to diagnose a disease is taken into account. Specificity is the measure of healthy patients being diagnosed as being healthy. Mathematically, it can be written as:

$$\begin{aligned} \text{specificity} &= \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}} \\ &= \frac{\text{number of true negatives}}{\text{total number of well individuals in population}} \\ &= \text{probability of a negative test given that the patient is well} \end{aligned} \tag{5.1}$$

A positive result in a test with high specificity helps in understanding whether the disease is present in the patients. The test rarely gives a positive result for healthy patients. Thus, a test with 100% specificity will rule out the disease from the healthy people accurately.

5.2 Sensitivity

Sensitivity is also a statistical measure used for calculating the performance of a classification test. Sensitivity is also called true positive rate. It is used to measure of actual positives which are computed properly for example percentage of healthy people being identified as healthy. Mathematically, it can be expressed as:

$$\begin{aligned}\text{sensitivity} &= \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}} \\ &= \frac{\text{number of true positives}}{\text{total number of sick individuals in population}} \\ &= \text{probability of a positive test, given that the patient is ill}\end{aligned} \tag{5.2}$$

Thus, a predictor with 100% sensitivity will recognize all the patients with the disease. Sensitivity is not the same as precision or positive predictive value. The calculation of the sensitivity does not take into account intermediate test results. If a test cannot be completed, then its intermediate samples can be excluded from the analysis or can be treated as false negative.

The skin lesion is classified according to the features obtained from the image. The optimal numbers of features are required to be obtained for good classification of the skin lesion. Selection of the features affects the specificity and the sensitivity of the classification. Thereby, with addition of each feature the effect of the feature to the performance of the classifier is noted. The trial and the error process are conducted on many features and 56 optimal features are obtained. These features provide specificity of 0.7692 and sensitivity of 0.973. One important observation was that the accuracy of the

classification did not depend on the number of features rather it depended on the type of features being added. Important features gave a better accuracy rate than the less important features.

5.3 Dependence of Performance of Classifier on Confusion Matrix

Confusion matrix is also called the error matrix. It is a table which visualizes the performance of an algorithm. Each column represents the instances in a predicted class while each row represents the instances in actual class. Confusion matrix is a table with two rows and two columns which represent the number of false positives, false negatives, true positives and true negatives.

In this work, initially three classes are taken such that the confusion matrix becomes 3*3 matrixes. Three rows represent the instances in actual class and three columns represent instances in predicted class. The specificity and sensitivity for 3*3 matrix is 0.7692 and 0.973 respectively. The numbers of classes are reduced to two such that the confusion matrix is 2*2. The specificity and sensitivity are 0.8461 and 0.973 respectively. With the increase in the classes better classification is observed. Therefore, the performance of the classifier depends on the number of classes taken into account. Further, if we incorporate a fourth class whereby the confusion matrix becomes 4*4, then the sensitivity and specificity changes to 0.8462 and 0.9189, respectively.

5.4 Support Vector Machine

A support vector machine is a classifier defined by a separating hyper plane. This algorithm provides the optimal hyper plane required to classify the labelled data in appropriate manner.

There can be a large number of hyper planes which can segregate the labelled data, but the optimal hyper plane is chosen. A hyper plane is not considered as optimal if it passes too close to the points. Hence, to find the optimal hyper plane it needs to pass as far as possible from all points. The SVM algorithm is based on finding the hyper plane that gives the largest minimum distance to the training examples. Twice, this distance is called margin. The optimal hyper plane maximizes the margin of the training data.

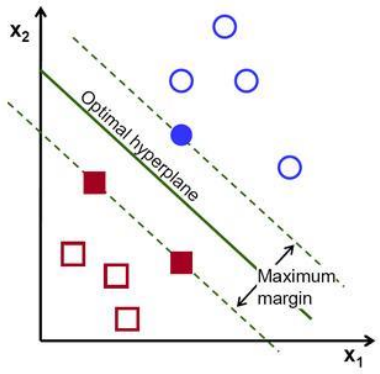

Figure 5.1 Hyper plane used for separating labelled data.

The equation for the optimal hyper plane is given by:

$$f(x) = \beta_0 + \beta^T x, \tag{5.3}$$

Where β is known as the weight and β_0 is known as the bias.

The optimal hyper plane is represented by:

$$|\beta_0 + \beta^T x| = 1 \tag{5.4}$$

Where, x denotes the training example closest to the hyper plane. In short the training examples near to the hyper plane are called support vectors. This representation is called canonical hyper plane.

The distance between a point x and a hyper plane (β,β_0) is obtained through geometry which is:

$$\text{distance} = \frac{|\beta_0 + \beta^T x|}{\|\beta\|}. \tag{5.5}$$

For a canonical hyper plane, the numerator is equal to one and the distance to the support vectors is:

$$\text{distance}_{\text{support vectors}} = \frac{|\beta_0 + \beta^T x|}{\|\beta\|} = \frac{1}{\|\beta\|}. \tag{5.6}$$

Support vectors are elements of the training example that would change the position of the hyper plane if it is removed. They are the most critical elements of training set.

The margin M is twice the distance of the support vectors:

$$M = \frac{2}{\|\beta\|} \tag{5.7}$$

Next, we maximize M by minimizing the function L (β) subject to some constraints. The constraints help in modeling the optimal hyper plane required to correctly classify the train data.

This is given by:

$$\min_{\beta,\beta_0} L(\beta) = \frac{1}{2}\|\beta\|^2 \text{ subject to } y_i(\beta^T x_i + \beta_0) \geq 1 \ \forall i, \tag{5.8}$$

Where y_i represents each of the label of the training example.

For two- dimension values, the separation can be done by a line and thus it is linear separable. Whereas, for higher dimension values the separation is done with the help of hyper plane. Sometimes features are non-linearly separable in the original space. Thus, to make the data linearly separable the data points are transformed to the feature space and thus the data are linearly separated.

For example, when the features are not linearly separable then if we transform it to higher space for example to a parabolic plane, the features get separated more easily.

5.5 K-means

K means clustering is popular for cluster analysis in data mining. k means clustering is useful for partitioning n observations into k clusters in which each observation belongs to the cluster with the nearest mean. For a set of n observations, k means clustering aims at partitioning of n observations into k clusters so as to minimize the within cluster sum of squares. In other words, the objective is given by:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \tag{5.8}$$

Where μ_i is the mean of the points in the cluster.

In k – means initially random cluster centroids are initialized. The shortest distance between the data point and the initial cluster center is calculated. If the data point is near to cluster center 1 instead of cluster center 2, then the data point is put in cluster 1 and the cluster center is updated. Again, the shortest distance of the data points is calculated with respect to the updated cluster centers. In this manner, all the data are clustered, respectively.

The centroid is updated by the formula:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{5.9}$$

Where x_j denotes the data present in the cluster. S_i is the cluster to which the data point belong. $S_i^{(t)}$ is the number of data points present in the cluster. $m_i^{(t+1)}$ is the updated cluster center.

5.6 Dependence of the Performance of the Classifier on SVM and K-means

SVM and K means clustering both can be used as classifiers. In this work, classification of the features where done by both SVM and K means and the performance that is the accuracy of the classifier was noted. The accuracy of the training examples when K means clustering is used is 90.3% while when SVM is used, the accuracy of the training examples increased to 95.6%. Thus, it is clear that SVM is a better classifier than the K means clustering. This is so because K means clustering works best for images that have the same variance that is the data points in the cluster are similar. Also, K means clustering has a problem of converging to a local optimal point whereas in SVM, the classification converges to the global optimal point and thus classification is far better.

CHAPTER 6

OPTIMAL CLUSTERING OF DATA

The data which is obtained contains three observations mild, severe and melanoma. Each observation has 56 features. These observations are clustered optimally with the help of genetic algorithm. In this work, the clustering is not done by k-means because k- means clustering normally converges to a local optimal solution whereas genetic algorithm provides a global optimal solution.

6.1 Optimal Clustering via Genetic Algorithm

The genetic algorithm is used for finding the optimal number of clusters. The procedure followed for obtaining the optimal number of clusters is given by the following flow chart:

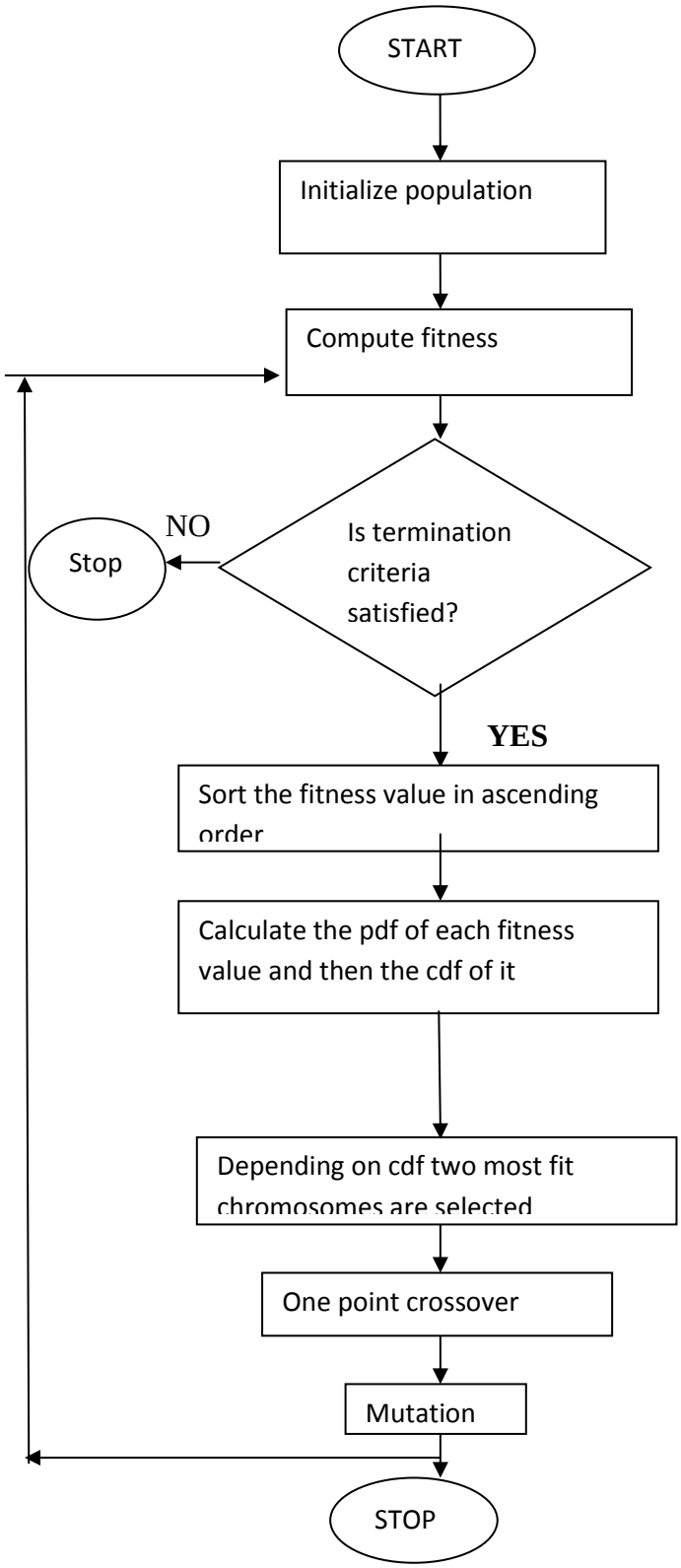

Figure 6.1 Flowchart for obtaining optimal numbers of clusters by genetic algorithm.

The numbers of clusters are initialized in the starting of the genetic algorithm. Each chromosome represents the cluster centers. The length of the chromosome is number of clusters by number of features in each observation. Suppose K cluster centers are present, then K clusters centers are encoded in each chromosome and are initialized to K random points in the Data set. This process is repeated for each of the chromosome in the population, where P is the size of the population.

The fitness computation process consists of two phases. In the first phase, the clusters are formed according to the centers encoded in the chromosome under consideration. This is done by assigning each observation x_i, where I = 1, 2 …..,n to one of the clusters C_j with the center such that:

$$\|\mathbf{x}_i - \mathbf{z}_j\| < \|\mathbf{x}_i - \mathbf{z}_p\|, p = 1, 2, \ldots, K, \text{ and } p \neq j. \tag{6.1}$$

After clustering is done, the cluster centers encoded in the chromosome are replaced by the mean points of the respective clusters. In other words, for clusterC_i, the new center z_i^* is computed as:

$$\mathbf{z}_i^* = \frac{1}{n_i} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j, \quad i = 1, 2, \ldots, K. \tag{6.2}$$

Then z_i^* now replace the previous z_i in the chromosome. Subsequently, the data inside the cluster has to be homogeneous that is the inter dissimilarity between the data points in the cluster has to be minimum.

The homogeneity in the cluster is calculated by the cluster matrix which is given by $\mathcal{M}$ and is computed as:

$$\mathcal{M} = \sum_{i=1}^{K} \mathcal{M}_i,$$

$$\mathcal{M}_i = \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \mathbf{z}_i\|. \tag{6.3}$$

The fitness function is given by $1/\mathcal{M}$ so that, maximization of the fitness function leads to minimization of $\mathcal{M}$.

The selection process selects the fittest chromosomes from the mating pool. Roulette wheel selection is the technique implemented. Roulette wheel selection is the common technique that implements the proportional selection strategy. The selection process selects chromosomes from the survival of the fittest concept. The fitness is first sorted and then the probability is calculated for each of the chromosome according to the fitness acquired by the chromosome. Cumulative distributive function (cdf) is computed for each of the chromosome and two fittest chromosomes are calculated.

Crossover is a probabilistic process that exchange information between two parent chromosomes for generating two child chromosomes. In this work, single point crossover with fixed crossover probability is used. Chromosomes of length l, a random integer called crossover point is generated in the range [1,l-1]. The portion of the chromosome lying to the right of the crossover point is exchanged to produce two offspring.

Each chromosome undergoes mutation with fixed probability. A bit position is mutated by considering a floating point. A number in the range of [0,1] is generated with uniform distribution.

The process of fitness computation, selection, crossover and mutation are executed for maximum number of iterations. The best string obtained in the last generation provides the solution to the clustering problem. The best string gives the optimal cluster points.

This is done for various numbers of clusters and the response of the Average fitness to the best fitness is observed. This trial and error process provides us six optimal cluster numbers.

The response of the genetic algorithm is given as:

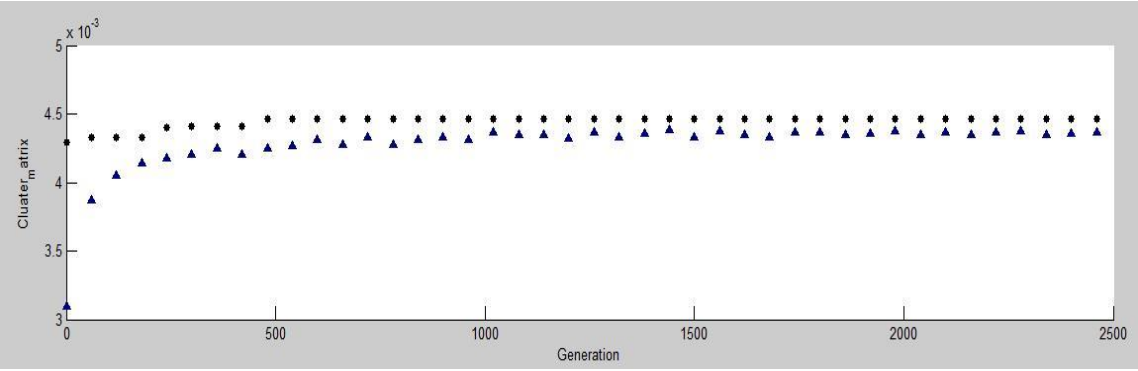

Figure 6.2 Optimization curve for the optimal number of clusters.

The crossover and mutation rate used are 0.7 and 0.3, respectively. A random number is chosen and if the random number is less than the crossover rate then the crossover operation is performed. In the same manner, when the chosen random number is less than the sum of the crossover rate and the mutation rate then mutation is performed. The crossover rate is first changed to 0.4. We observe that the optimal value is reached very fast thus, it can be concluded that when the crossover rate is decreased, the solution can converge to local optimal solution. In the same way, when the crossover rate is increased to 0.9, we see that the optimal solution takes a longer time to be obtained. Therefore, it is clear that to get a global optimal solution the crossover rate should be in the mid-way.

CHAPTER 7

CLASSIFICATION OF CLUSTERS

The optimal numbers of cluster are obtained from genetic algorithm. The main criteria for classification of the clusters are that the data in each of the cluster has to be homogeneous.

The clusters centroids are obtained from genetic algorithm. The best chromosome contains the optimal cluster centers which are obtained from the fitness function. The fitness function is obtained by minimizing the inter dissimilarity inside each cluster. Thus the optimal clusters contain data which is homogeneous in each cluster. These clusters are labelled and homogeneous such that all members in a cluster belong to the same class.

Each cluster is then given to an SVM prototype. The flow chart of the SVM Classification is:

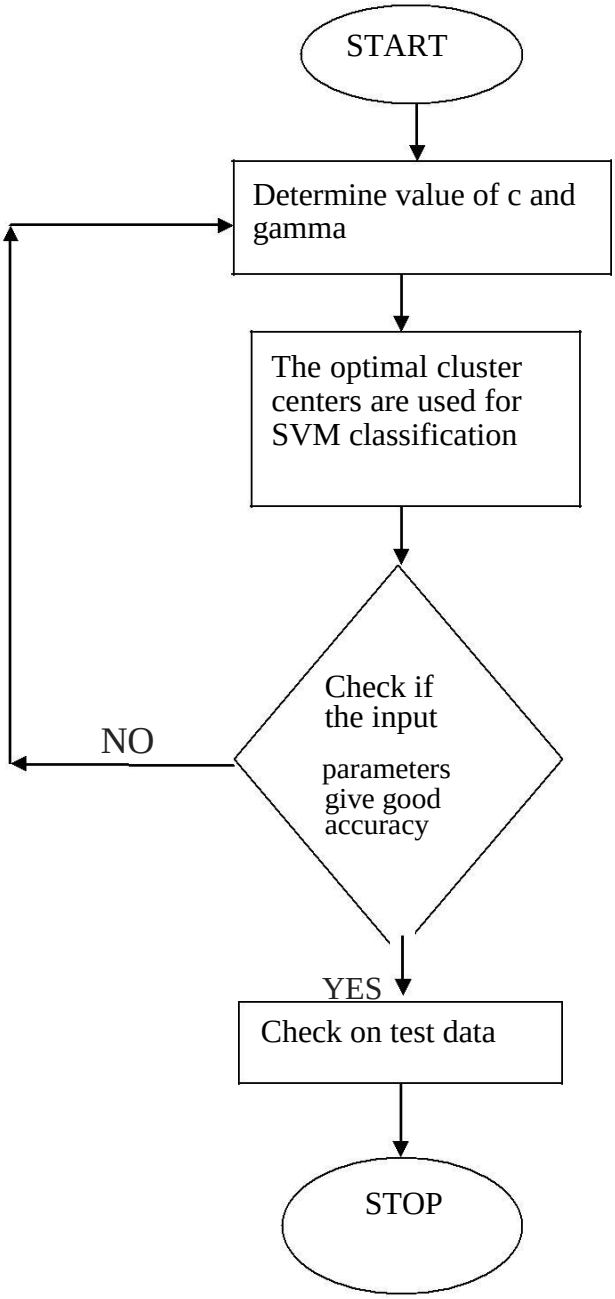

Figure 7.1 Flowchart for classification of clusters.

The clusters are transformed to a higher space so that they get linearly separated in the feature space. The feature space is determined by the kernel function. There are a lot of kernel functions like radial base function, linear, polynomial etc.

The choice of the kernel is a trial and error process. For each kernel which is used, specificity and sensitivity is obtained. It is observed that the radial base function provides the best specificity and sensitivity values. The observations show that the radial base function is the best kernel which can be used. The radial base function is a real valued function whose value depends on the distance from the origin or on the distance from any value C. Any function that satisfies the property is a radial function. The distance which is being used is a euclidean distance although, there can be other distances possible as well. In the radial function, C and gamma parameters are used. C is the penalty factor. This parameter needs to be chosen properly because if the value of C is very large then there is a chance of over fitting whereas, if the value of C is small, chances are there of under fitting. The parameter C also controls the trade-off between the errors of the SVM on training data and the margin maximization. Gamma is used to obtain the shape of the separating hyper plane. Increasing gamma usually increases the number of support vectors.

In this work, the gamma and C values are taken as 10^-9 and 10^3. The clustered data for SVM classification increases the response speed. The response speed is more than the SVM classifiers. Moreover, the testing accuracy is more and can be guaranteed to quite some extent.

CHAPTER 8

OPTIMAL RADIAL BASIS PARAMETERS AND FEATURES

The SVM is used to classify the data points in each cluster. The parameters used in the kernel needs to be optimized for obtaining a better accuracy rate. To minimize this error, again genetic algorithm is used. Genetic algorithm not only minimizes the error but also helps in obtaining the optimal parameters for radial basis function and also the optimal features.

The flowchart for the genetic algorithm is:

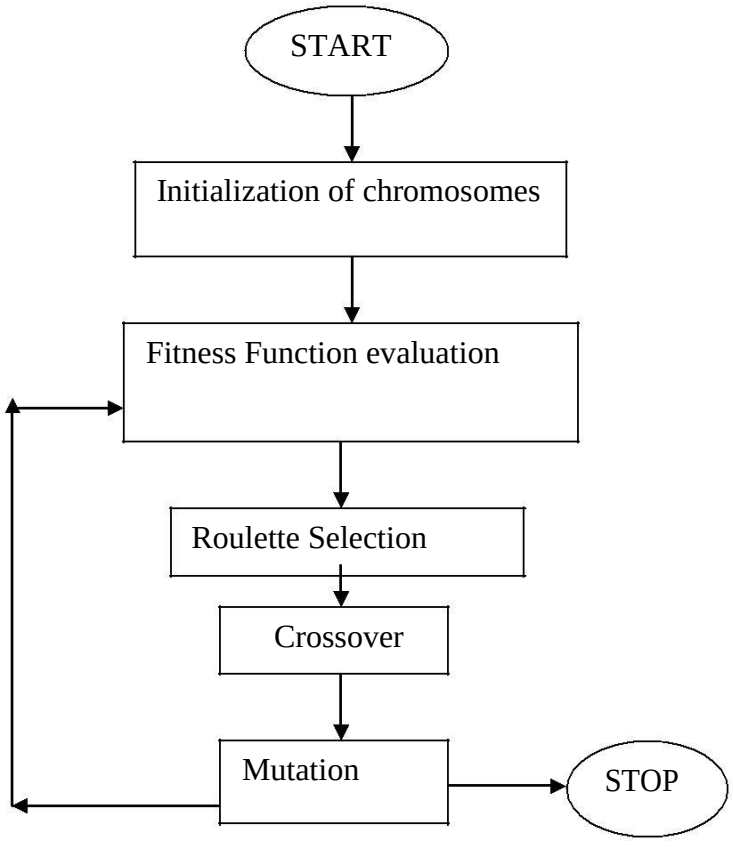

Figure 8.1 Flowchart to obtain optimal radial base parameters.

When the RBF kernel is selected, the parameters and features were used as the input attributes in Genetic algorithm. The chromosome comprises of three parts, C, ϒ, number of clusters and the feature mask. However, these chromosomes have different parameters when other types of kernel functions are selected. Classification accuracy and the slope of the SVM error are the criteria used to design a fitness function. The fitness function consist of two predefined weights W_a and W_b . These two weights are normalized such that W_a + W_b =1. The SVM error plot is obtained and then the slope is calculated for 2500 iterations:

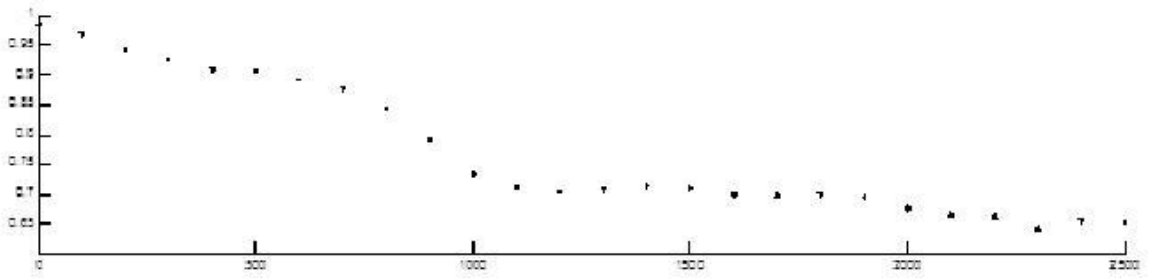

Figure 8.2 SVM error plot.

A higher value of weight accuracy W_a is used for obtaining higher accuracy value. The number of iterations is taken as 500.Thus the fitness function is:

$$\text{Fitness} = W_a * \text{SVM\_accuracy} + W_b * \text{slope} \qquad (7.1)$$

SVM_ accuracy value is obtained from the SVM. The main advantage of scaling is to avoid attributes in greater numeric ranges.

Another advantage is to avoid the numerical difficulties during the calculation. Feature value scaling help to increase the SVM accuracy. The training dataset is used to train the SVM classifier while the testing dataset is used to calculate classification accuracy. When the classification accuracy is obtained, each chromosome is evaluated by fitness function. The termination criterion is the number of generation. When the termination criterion is reached, the process ends. In this work W_a is 0.82 and W_b is 0.18.

The fitness value is associated to each chromosome in the population. After the fitness value is obtained, two best chromosomes with the maximum fitness value are obtained. The two best chromosomes are selected by first finding the pdf of each chromosome and then sorting the pdf. Cumulative distributive function is obtained for each sorted chromosome. A random number is taken and the sum of the cumulative distributive function less than the random number delivers the indexes of two chromosomes which is selected for crossover and mutation process. The crossover is done on two selected chromosomes. In this work, uniform crossover is conducted. Mutation is also performed.

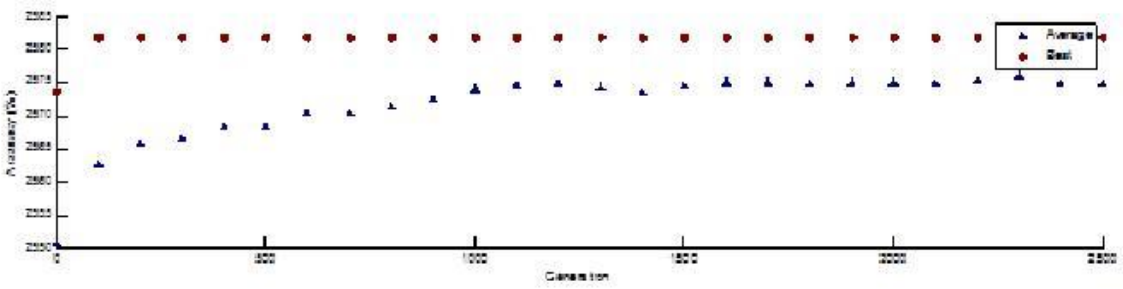

Figure 8.3 Optimization of Radial bases parameters.

.

CHAPTER 9

RESULTS

The original image which is obtained initially contained artifacts. Initially, the artifacts were removed and the image was made artifact free. Then the texture and the color features where obtained from the lesion. This work contains three sub parts. First, the optimal number of features where obtained which could give the maximum amount of specificity, sensitivity and accuracy. Second, the optimal numbers of clusters are obtained. Third, the optimal parameters for radial base function are obtained so that optimal parameters along with the optimal features are used to obtain a high accuracy rate of the SVM classifier.

9.1 Results for Obtaining Optimal Features

A total of 56 optimal features are obtained. These features are obtained by adding each feature and observing the change in the specificity, sensitivity and accuracy of the SVM.

The observation is as follows:

Table 9.1 Specificity, Sensitivity and Accuracy Rate for Different features

Features used	**Number of Features**	**Accuracy %**	**Specificity**	**Sensitivity**
GLCM offset-1,2,3,4,5 and wavelet – 3 levels	38	90%	0.973	0.6154
GLCM offset – 1,2,3,4,5 and wavelet – 3 levels(dimension is increased to 20)	48	90%	0.999	0.7692
Wavelet – 3 levels(dimension decreased to 16) and GLCM offset- 1,2,3,4,5	58	94%	0.973	0.8461
Wavelet- 3 levels(dimension increased to 25), GLCM offset-1,2,3,4,5, added feature ASM	57	90%	0.973	0.692
Wavelet – 3 levels (dimension is 20), GLCM offset-1,2,3,4,5, added features ASM, standard deviation, entropy, variance	64	94%	0.981	0.7692
Wavelet – 3 levels(dimension is 16), GLCM offset- 1,2,3,4,5, added features ASM, standard deviation, entropy	56	96%	0.991	0.927

The observation shows that with increase in the number of features, the SVM accuracy may or may not increase. Whereby, the sensitivity and specificity does not depend on the number of features. The observation table shows that 56 optimal features give the best accuracy percentage, specificity and sensitivity.

The sensitivity and the specificity of the classifier also depend on other parameters which are being used such as the SVM classifier rather than K-means, addition of new class and many more. Each of the parameters is changed and their effects on sensitivity, specificity and accuracy rate are noted.

9.2 Relation of Sensitivity and Specificity with Respect to Confusion Matrix

Parameters like confusion matrix are used to obtain the specificity and sensitivity of the classifier. The observation table is obtained for the confusion matrix:

Table 9.2 Sensitivity, Sensitivity and Accuracy Rate for Different Confusion Matrix

Parameters	**Number of features**	**Accuracy %**	**Specificity**	**Sensitivity**
Confusion matrix as 2*2	56	98%	1	0.9231
Confusion matrix as 3*3	56	96%	0.991	0.927

From the observation table, it can be very well concluded that the sensitivity and the specificity of the classifier depends largely on the confusion matrix. As the size of the confusion matrix increases, the sensitivity of the classifier also increases. When the size of the confusion matrix is 2*2, it means that there are 2 classes whereas, when the size of the confusion matrix is 3*3, it means that there are 3 classes. It is observed that with the

Increase in the classes the accuracy rate, specificity and sensitivity of the SVM classifier changes. Thus, it is clear that specificity, sensitivity and accuracy rate depends on the number of classes being used.

9.3 Types of Classifier used

The performance of the classifier also depends on the type of classifier being used. SVM and K-means are the two types of classifiers that are being used. Observations are obtained for SVM and K means classifiers:

Table 9.3 Type of classifier used

Parameters	Number of features	Accuracy rate	Specificity	Sensitivity
SVM	56	95.60%	0.9	1
K means	56	91.30%	0.9	0.9231

The observation makes it very clear that SVM is a better classifier than K means. This is so because SVM does not get stuck in local centroids and thus provides better accuracy rate than K means.

The next step was to obtain the optimal cluster numbers. The cluster numbers were obtained by genetic algorithm. The cluster numbers were obtained by observing the plot of the best vs. the average fitness convergence rate.

By observation, it was found that if the total number of clusters were taken as 6, then it gave the best result and the convergence occurred near to 1000 generations

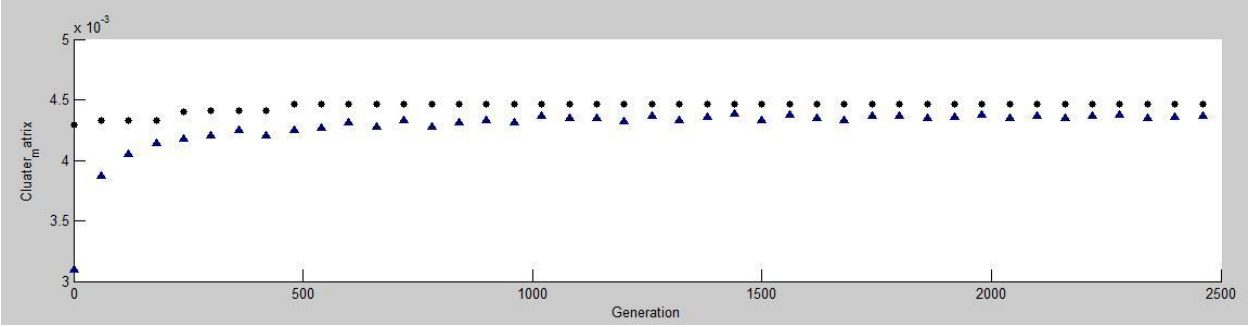

Figure 9.4 Plot for optimal clusters.

The next step was to classify the features in each cluster. Features in each cluster are classified with the help of SVM.

9.4 Choice of kernel

The choice of the kernel for the SVM is done by using each kernel and observing the specificity and sensitivity offered by it during classification. The observation is as follows:

Table 9.4 Relation of sensitivity and specificity for the choice of kernel

Kernel Used	**Specificity**	**Sensitivity**
Linear	0.6250	0.9091
Polynomial	0.8750	0.3636
Radial base function	0.9816	0.9238

The observation shows that the radial base function provides the best specificity and sensitivity. It is clear that radial base function is the best choice for the kernel which is being used.

The last step of the work was to find the optimal values of the parameters for the radial base function. The optimal values were obtained from the fitness value of the genetic algorithm. Fitness value required the SVM error plot and the values of W_a and W_b . The values of W_a and W_b are taken as 0.82 and 0.18 respectively. The genetic algorithm provide the optimal radial base parameters and the optimal features. It provides a test accuracy of 90% and a training accuracy of 96.42% and also the fitness value which is obtained is 90.

Thus, the plot which is obtained for the optimal parameters of radial base function and features is:

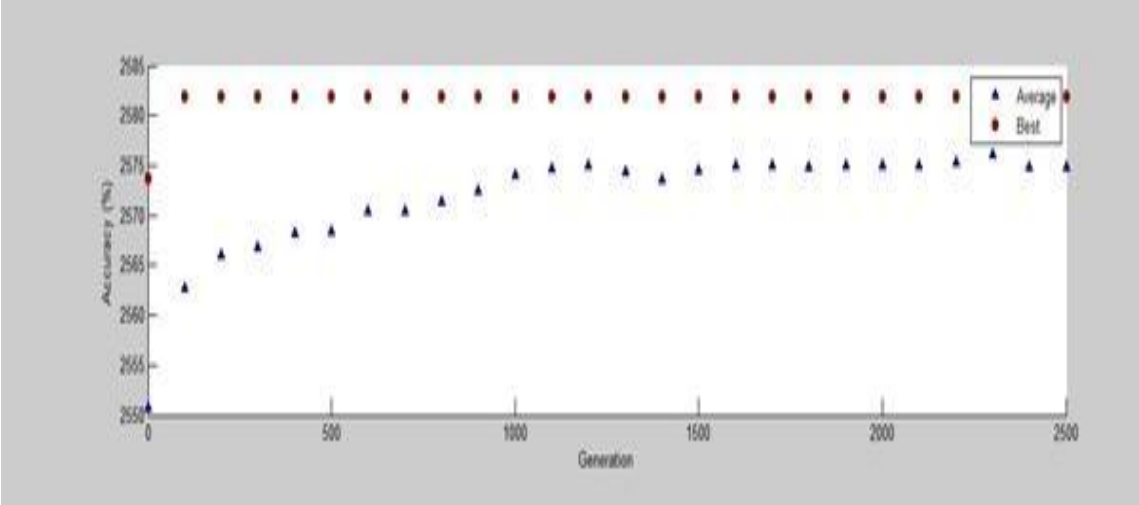

Figure 9.6 Plot of optimized kernel parameters and feature set.

The training accuracy and the test accuracy which is obtained by this work is far better result than the result which was obtained before.

Therefore, by selecting the optimal features and clusters and then classifying the homogeneous data with optimal kernel parameters help in getting a better result for test and training accuracy.

CHAPTER 10

CONCLUSION

In this work, a lot of methods and investigations are done to improve the accuracy of the skin cancer early. Such methods could be instrumental for the quick and inexpensive screening of suspected skin lesions to aid dermatologist in the diagnosis.

The two dimensional texture analyses of the TLM and ELM images helped in classifying the skin lesions in three layers of dysplasia. However, the accuracy of this analysis is limited by the fact that white illumination was used. This demonstrates the need for multispectral imaging as a better way to measure the subsurface blood volume.

The first contribution to the work was to get the optimal features. The optimal features are selected and the features help in obtaining a better accuracy value. The features which are obtained in this work are mostly color and texture features. The color and the texture features are the most important features used for analyzing the skin lesion. There are also other features which can be used for future work like the features obtained from fluorescence intensities, kinetics of the skin and many more. Though these features provide significantly less amount of effect in analyzing the lesion, they can be taken into account for better obtaining better level of accuracy.

The next step was to find the optimal clusters for obtaining better accuracy of the skin lesion. The optimal clusters help in clustering the data and thereby speeding up the classification process. Clustering is an important unsupervised classification where a set of patterns, usually vectors in a multi-dimensional space are grouped into clusters into such a way that the patterns in different clusters are dis-similar in the same sense. For this

it is necessary to first define a measure of similarity which will establish a rule for assigning patterns to the domain of a particular cluster center. One such measure of similarity may be Euclidean distance. Smaller the distance between the data point and the cluster center, greater is the similarity between the two. An intuitively simple and effective clustering technique is the well-known K means algorithm. However, it is also known that the K-means algorithm may get stuck to suboptimal solutions depending on the choice of the initial cluster centers. In this work, a solution to the clustering problem where genetic algorithms are used for searching the appropriate cluster centers such that a given data set is optimized.

Genetic algorithm is a random search and the optimization search is guided by the principles of evolution and natural genetics, and having a large amount of implicit parallelism. Genetic Algorithm perform search in a complex, large and multimodal landscapes and provide near optimal solutions for the objective or fitness function of an optimization problem. Therefore, under limiting conditions, a genetic algorithm based clustering technique provides an optimal clustering with respect to the clustering data set being considered.

The next interesting contribution was to obtain the optimal parameters for radial base function. In this section, SVM parameters and the features subsets were optimized simultaneously because the selected feature subset has an appropriate kernel parameters and vice-versa. A genetic algorithm based selection strategy was made to select the feature subset and the set of optimal parameters for SVM classification.

In this work, parameters of radial base function kernel are optimized. However, other kernel parameters can also be optimized by using the same approach. The proposed

algorithm can also be applied to support vector regression. The kernel parameters and the input features heavily influence the predictive accuracy of the support vector regression with different kernel functions. The same genetic algorithm based feature selection and parameter optimization procedures to improve the SVR accuracy. The information gained from the Trans illumination imaging and the proposed methods are useful as indicators of early malignancy, such as increased angiogenesis and lesion severity. These tools and methods would be useful to a dermatologist as additional information in the decision of the biopsy. Most importantly, this approach to optical imaging could be used as a quick and inexpensive method for mass screening of the patients for early detection and diagnosis of skin cancers. Lesions could be tracked over time to detect physiological changes indicative of early malignancy.

Since survival is quite high for the patients who have malignant lesions detected at an early age, improved tools for accomplishing this early detection are necessary. Furthermore, the potential impact of these methods is far reaching in the fields of non-invasive diffuse optical imaging, not only the characterization of skin lesions to facilitate early detection of skin cancer, but also for other applications of tissue characterization such as neural imaging.

Our proposed algorithm shows much promise in the ability to classify the different grades of skin lesion dysplasia. Our TLM background and color correction algorithm along with our lesion segmentation algorithm and interactive interface are clearly able to highlight the increase in vascularity present in increasingly dysplastic lesions. The trend we observe in the TLM/ELM ratio is highly significant.

Classification of lesions using SVM shows good promise in grading the specific severity of lesion dysplasia, with an aim towards grouping lesions into classes where appropriate action can be taken by a dermatologist. These tools and methods would be useful to such a dermatologist as additional information to assist in the decision to biopsy. Our goal in the future is to improve our classification results through the intelligent selection of other distinguishing features using multispectral imaging to reliably classify our three levels of dysplastic skin lesions

APPENDIX A

WAVELETS

A.1 Time- Frequency Analysis

The Heisenberg uncertainty principle states that it is impossible to the exact location and momentum of a particle at the same time. The measurement of one property becomes more precise, measurement of the other property losses its precision. The uncertainty principle also has application in time-frequency analysis of signals where it is impossible to localize a point in both the dimensions.

A given continuous signal in the time domain has perfect time resolution. An exact amplitude is known for every specific time that is desired. However, nothing is found about the frequencies contained in the signals. For frequency analysis, Fourier transform is employed and the purpose of it is to decompose a time domain signal into a frequency domain signal to find the frequencies present in the signal. The transform is found by integrating over the entire length of the original signal from negative infinity to positive infinity.as:

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-j\omega t} dt \qquad \text{(A.1.1)}$$

Where f(t) is the signal in time domain, and f(w) is the signal in frequency domain. With the Fourier transform, frequency information is obtained by losing all the information about time.

The complexities increases by splitting the time domain signal into multiple sections and taking Fourier transform of each section. By observing the frequency domain of each section, one can still determine the frequencies that are present. The accuracy or the resolution can be improved by the number of sections that are made in the original signal. However, one cannot keep increasing the sections in the original signal, the uncertainty principle come into action. As more divisions are created, the frequency analysis of each section becomes less accurate. Thus, a balance is required between the temporal resolution and the frequency resolution for the signal analysis.

A method of sectioning the time frequency plane into separate areas is known as short-time Fourier transform. The division of time into discrete chunks is most commonly done by using a window function. This window function has compact support and is shifted and multiplied with the signal along with the time axis to produce the divisions in time needed for frequency analysis of each division.

A.2 Continuous Wavelet Transform

The short-time Fourier transform is good tool for multi resolution analysis of signals, but is very dependent on the size of the window chosen. A fixed window often results in poor frequency resolution in the low frequency range and poor time frequency resolution in high frequency range. To overcome this difficulty, wavelet analysis was developed. It is a versatile tool which provides high frequency resolution where time resolution is not important and low frequency resolution where time resolution is important. The general idea of accomplishing this is by shifting and scaling a window across the signal to produce multi - resolution decomposition.

A wavelet is essentially a function that meets certain criteria such as having finite energy, zero frequency component resulting in the wavelet having a band pass spectrum. The spectrum is given by:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right) \tag{A.2.1}$$

The weighting constant in front of the wavelet function ensures that wavelets at all scales have same energy. The wavelet transform simply involves projecting the signal onto the bases produced by the shifted and scaled version of the original wavelet:

$$\begin{aligned} w(u,s) &= \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\psi^*\left(\frac{t-u}{s}\right)dt \\ &= \int_{-\infty}^{+\infty} f(t)\psi^*_{u,s}(t)dt \end{aligned} \tag{A.2.2}$$

In a sense, these replace the concepts of time and frequency in the Fourier transform. The translational shift becomes a measure of time resolution, while scale, or width, is what defines frequency resolution.

APPENDIX B

GENETIC ALGORITHM

B.1 Initialization

Initially all the solutions are randomly generated to form the initial population. The population size depends on the nature of the problem and contains as many solutions as possible. Generally the solutions are generated randomly allowing the entire range of possible solutions. Individuals in the population represent the different solutions for the problem. Information of these individuals are encoded in the chromosome.

B.2 Evaluation of the Fitness

The fitness value is obtained from the problem which is being dealt with. It describes the ability to reproduce and survive. Fitness is a probability rather than an actual number of offspring. The fitness is the probability which helps in identifying whether the individual will be included among the group selected as parents or the next generation. The fitness of an individual is modeled through the use of fitness function. The fitness function takes a single individual or a single chromosome as an input and evaluates how well the problem is solved with the help of the input chromosome. The fitness function returns the fitness evaluated. Chromosomes with higher fitness have a greater chance to reproduce and pass their information to the offspring. With each generation, the population of the chromosome should converge closer to the optimal solution of the problem.

A chromosome is most often represented as a one dimensional binary string. The chromosome can be either binary or a set of real valued numbers. The actual shape of the chromosome may vary as well. Regardless of the actual format of the chromosome, the only requirements are that it contains finite set of numbers and it can evaluate the fitness function.

B.3 Selection

During each generation, a proportion of the existing group is selected to reproduce the next generation. The proportion of the existing group which is selected to reproduce the next generation is done with the help of the selection process. Individual solutions are selected based on the fitness value. Certain selection processes rank the fitness of each solution and preferentially select the best two. One of the selection process is the roulette wheel selection. In this process each chromosome has a small chance of being selected but the chromosome with higher fitness value has a higher chance of being selected, A simple way of finding the selection probability p_j for each chromosome in the population size N with fitness evaluation [F_1,F_2 ,…..,F_N] is:

$$p_j = \frac{F_j}{\sum_{i=1}^{N} F_i} \qquad \text{(B.3.1)}$$

However, this technique does not perform well in populations where the fitness evaluations have a small percentage deviation from each other. Alternatively, a fitness technique known as linear normalization scales the fitness values to a set range of

[0, 100] before computing the selection probability. Another method ranks the fitness values in the ascending order. The selection probability of the chromosome is:

$$p_j = \frac{\text{rank}(F_j)}{0.5 \cdot N(N+1)} \tag{B.3.2}$$

This method sorts the probability in ascending order with the least fit to most fit chromosome.

B.4 Genetic Operators

The next step is to generate the second generation of population from those selected through a combination of genetic operators: crossover and mutation.

Each new solution to be produced, a pair of parent solutions is selected for reproduction from the pool selected previously. A child solution is obtained from the crossover and mutation which is created from the parent solutions. New parents are selected for each new child and the process continues until a new population of solutions of appropriate size is not generated. These processes ultimately generate the next generation of population which is different from the initial population and the average fitness will be increased for the next generation. There are a number of reproduction techniques which includes generational replacement and steady-state reproduction. In generational replacement, the entire population is replaced by the offspring in each generation while in steady –state reproduction one or two chromosomes are replaced by their offspring in each generation. Thus, it is clear that steady –state reproduction requires larger time to converge than generational replacement. However, when we compare the

performance of the two methods, steady-state reproduction produces better results than generational reproduction.

B.5 Crossover

Once the selection probability of each chromosome is obtained, two parents are picked based on the two best probabilities. These two best chromosomes are selected for reproduction. Reproduction is done by "crossover". Crossover is a method by which two parent chromosomes are mixed to produce two new offspring in their place. The simplest method of cross over is one point crossover. Each parent that is chromosome A and chromosome B are divided into two sections.

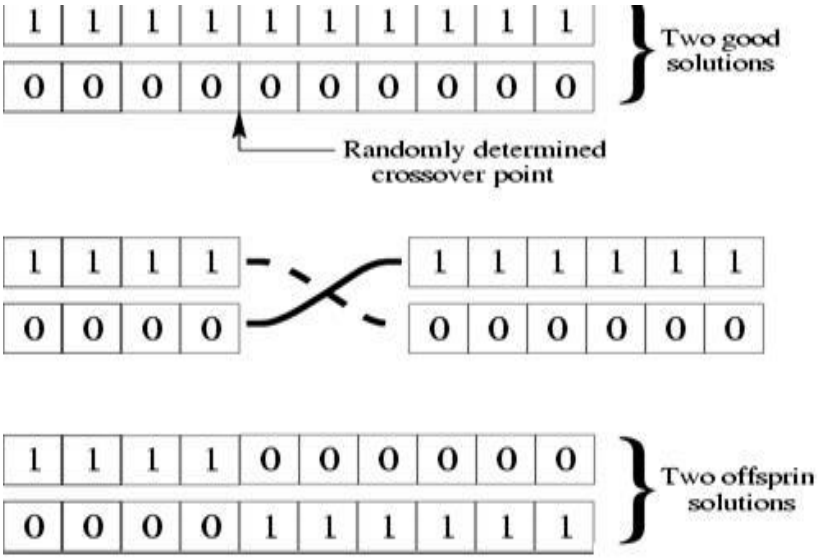

Figure B.5.1 Single point Crossover.

Chromosome A is [1 1 1 1 1 1 1 1 1 1] and Chromosome B is [0 0 0 0 0 0 0 0 0 0]. This division is often in the middle of the chromosome although the position can be varied randomly. The first section of the Chromosome A is attached to the second section of the

Chromosome B. This is the first offspring. Likewise, the second part of the Chromosome A becomes attached to the first section of the Chromosome B. This forms the second offspring. These two offspring is placed into the pool for next generation.

Another method of crossover is known as the uniform cross over. Below figure B.5.2 for uniform crossover.

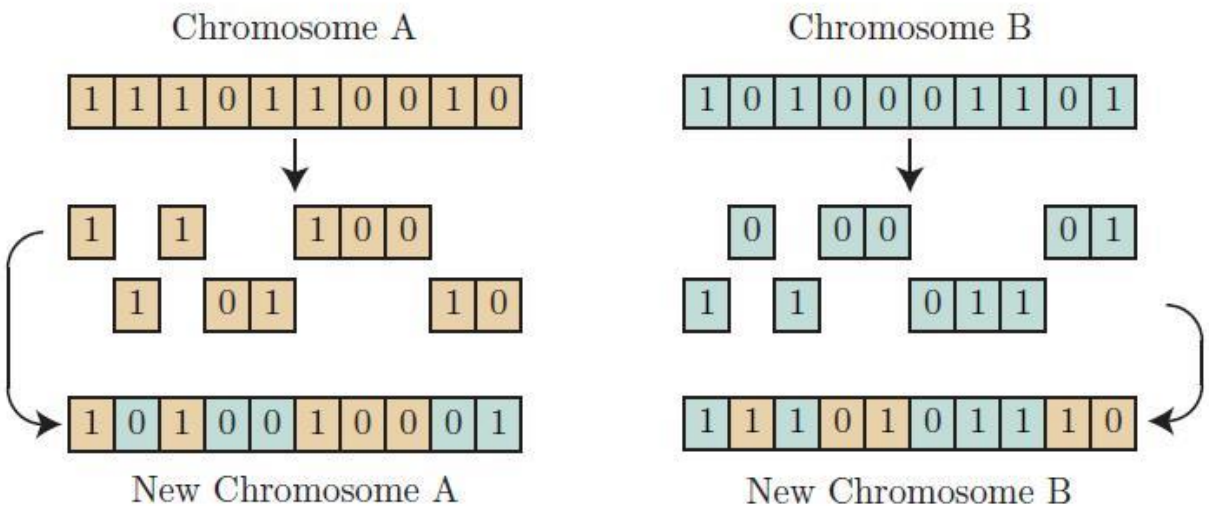

Figure B.5.2 Uniform Crossover with mask [1010011100].

In this method, a binary mask is randomly generated for each of the two parents which divide each of the parent chromosomes into many sections. To produce the first offspring, positions where the mask is equal to 1 take their value from Chromosome A while where the value of the mask is equal to 0 take the value from Chromosome B. This gives the first offspring while the opposite procedure produces the second offspring.

B.6 Mutation

Another genetic operator is Mutation. Within each generational loop, each chromosome in the population has the probability of randomly changing one of its values. For binary chromosomes, one bit is randomly flipped. For real valued chromosome, a new random number can be replaced on one of the positions in the chromosome. Mutation adds a new generic element into the chromosome. This helps in converging to the global minima as mutation allows the population to explore small changes around the solution to find the optimal solution within the given solution neighborhood.

REFERENCES

1. Brian D'Alessandro, Atam P. Dhawan and Nizar Mullani," Computer Aided Analysis of Epi-illumination and Transillumination Images of Skin Lesions for Diagnosis of Skin Cancers," in 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, August 30 - September 3, 2011.

2. Cheng-Lung Huang and Chieh-Jen Wang,"A GA based feature selection and parameters optimization for support vector machines," Expert systems with applications, vol.31, pp.231-240,2006.

3. Ujjwal Maulik and Sanghamitra Bandyopadhyay," Genetic algorithm-based clustering technique," The journal of the pattern recognition society, vol.33, pp.1455-1465, April 1999.

4. Mamta Mor, Poonam Gupta, Priyanka Sharma, "A Genetic Algorithm Approach for clustering," International journal of engineering and computer science, vol.3, pp.6442-6447,6th June 2014.

5. Bashar Al-Shboul, Sung-Hyon Myaeng , "Initializing K-Means using Genetic Algorithms," World Academy of Science, Engineering and Technology, pp.54, 2009.

6. Melanie Mitchell, An Introduction to Genetic Algorithms (5th edition). [On-line].Available: http://www.boente.eti.br/fuzzy/ebook-fuzzy-mitchell.pdf [April,2014].

7. L. Xua, M. Jackowski, A. Goshtasby, D. Roseman, S. Bines, C. Yu, A. Dhawan, A. Huntley:" Segmentation of skin cancer images," Image and vision computing , vol.17, pp:65-74, 1997.

8. Jiaqi Wang, Xingdong Wu, Chengqi Zhang ,"Support vector machines based on K-means clustering for real time business intelligent systems," Int. J. Business Intelligence and Data Mining, vol. 1, pp.1,2005.

9. Abdul Ghaaliq Lalkher, Anthony McCluskey," Clinical tests: sensitivity and specificity," Contin Educ Anaesth Critic Care Pain, vol.8(6), pp.221-223,2008.

10. John Sikorsk,"Identification of malignant melanoma by wavelet analysis," in Proceedings of Student/Faculty research day, CSIS, Pace University, May 7th ,2004.

11. Nima Fassihi, Jamshid Shanbehzadeh, Abdolhossein Sarafzadeh, Elham Ghasemi "Melanoma Diagnosis by the Use of Wavelet Analysis based on Morphological Operators," in Proceedings of International Multi Conference of Engineers and Computer Scientists 2011, vol.1, IMECS 2011, March 16-18 2011, Hong- Kong.

12. Ammara Masood and Adel Ali Al-Jumaily," Computer Aided Diagnostic Support System for Skin Cancer: A Review of Techniques and Algorithms," International journal of Biomedical Imaging, article id: 323268, pp: 22 pages, 2013.

13. Chun-Hung Cheng, Wing-Kin Lee, and Kam-Fai Wong," A Genetic Algorithm-Based Clustering Approach for Database Partitioning ," IEEE Transactions of Systems, Man and Cybernetics- Part C: Applications and Reviews, vol.32, August 2002.

14. D.N.V.S.L.S. Indira, Jyotisna Supriya P," Detection & Analysis of Skin Cancer using Wavelet Techniques," International Journal of computer science and information technologies, vol. 2(5), pp.1927-1932, 2011.

15. Jayapriya J, D. Palanikkumar," Combining Supervised Attribute Clustering And Ga-Svm Classifier For Microarray Sample Classification," International Journal of Engineering Research and Technologies, vol. 2, June- 2013.

16. Wei Lu and Issa Traore," A New Evolutionary Algorithm for Determining the Optimal Number of Clusters "PO Box 3055 STN CSC, Victoria, B.C., Canada.

17. Christopher J.C. Burges," A tutorial on Support vector machines for pattern Recognition," Data Mining and Knowledge Discovery, vol.2, pp. 121-167, 1998.

18. I.N. Kapur, P.K.Sahoo , A.K.C. Wong, " A New Method for Gray-Level Picture Thresholding Using the Entropy of the Histogram," Computer vision,Graphics and Image Processing, vol.29, 273-285, 1985.

.

19. "RBF SVM parameters," http://www.tomzap.com/notes/TechCommunicationsEE333T/IEEE_ReferenceEx amples.pdf,, September 2014.

20. P. Mohanaiah, P. Sathyanarayana, L. GuruKumar, "Image texture feature Extraction using GLCM approach," International journal of Science and Research Publications, vol.3, Issue 5, May 2013.