Fall 2013

Feature extraction through k-means segmentation for melanoma detection

Snigdha Priya Bommadevara

Department:

FEATURE EXTRACTION THROUGH K-MEANS SEGMENTATION FOR MELANOMA DETECTION

by

SNIGDHA PRIYA BOMMADEVARA

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

2013

Approved by

Randy H. Moss, Advisor

R. Joe Stanley

William V. Stoecker

ABSTRACT

Malignant melanoma is responsible for 75% of the deaths caused due to skin cancer annually [1]. However, melanoma detection can be possible through feature extraction and pattern classification, which can lower the risk, if the melanoma is detected at an early stage. Clustering is one of the most useful tools used to differentiate features that can contribute to melanoma. This research work uses the k-means clustering algorithm for implementation of color segmentation. However, k-means clustering requires a predefined value of k, i.e., the number of clusters must be specified at the beginning of the run. This research uses a predefined value of k=4 determined empirically after numerous test runs on the data set. A set of 888 dermoscopy skin lesion images were used in this work with k-means segmentation used to segment the lesion area in each image into four colors, and 226 features were extracted from each image. Forward stepwise logistic regression (as implemented in the Statistical Analysis Software (SAS) package) was used for feature selection and model building. SAS returned 90 significant features and created a model with a diagnostic accuracy as measured by the area under the Receiver Operating Characteristic (ROC) curve of 0.902.

ACKNOWLEDGEMENTS

I would like to express deepest gratitude to my advisor, Dr. Randy Moss, for all his guidance, comments, contributions and valuable suggestions for this research work. Also, I would like to extend my gratefulness to my committee members, Dr. William V. Stoecker and Dr. Joe Stanley, who with their suggestions and encouragement have helped me throughout the course of the research. I feel privileged to have worked with all three of them and this study would not be possible without the help and support of the committee.

I would take this moment to express my gratefulness and thanks to my fellow members of the research group, Jason Hagerty and Nabin Mishra, who have constantly helped me with their invaluable inputs and contribution to the research work. Also thanks to the whole research group who in their own way contributed to the research. My special thanks to all my professors who taught me various aspects and nuances of Electrical Engineering during the whole tenure of Master`s degree and have always provided me support to move forward.

Last but not least I would like to offer special thanks to my parents, family, friends and the Almighty god for all the blessings and well wishes bestowed upon me.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Page

LIST OF TABLES

Page

1. INTRODUCTION

Malignant melanoma had an estimated incidence of 76,250 cases and 12,190 deaths in the United States in 2012 [2].Detecting these cancers at an early stage has a direct effect on the likelihood of survival. Melanoma in situ, the earliest stage of melanoma, does not affect life expectancy [3]. Features extracted with color as the basis can be useful in detecting the cancers thereby reducing the severity of the disease which might even lead to death in some cases.

Feature extraction in particular is very helpful in distinguishing malignant and benign lesions. Several methods can be used to extract features for classifying skin lesions. This work deals with using clustering to segment the image on the basis of colors.

1.1 K-MEANS ALGORITHM

The k-means algorithm partitions the given dataset into k predefined clusters using squared Euclidean distance [4]. Each object is considered as an n-dimensional vector where n is the number of features used as basis for the clustering. The algorithm assigns k centers initially using the Forgy method which serve as the centroids of the clusters, and the objects are assigned to the nearest cluster. When all the objects have been assigned to one of the clusters, the centroids are recalculated and the objects are now assigned according to the new centroids. This procedure is performed iteratively until there are no more changes.

1.2 DATA SET

This k-means segmentation was applied to a data set consisting of 888 contact non-polarized dermoscopy images taken using the 3Gen DermLite Fluid dermatoscope (3Gen Inc., Dana Point, CA). The image set was collected from patients in the course of the study NIH CA153927-02A2 from four private-practice clinics in Columbia, MO, Plantation, FL, Rolla, MO, and Stamford, CT. This set consisted of 195 melanoma images and 693 benign images.

Using this data, k-means segmentation separated the lesion area of each image into four segments taking the RGB value as the basis for segmentation. The lesion masks are overlaid onto the image to separate the lesion area from the background. These lesion masks used for overlaying were obtained by manually marking lesion areas with software (Winshow) based on a second-order spline technique, and the borders were verified by a dermatologist. Winshow generates a ternary mask which has a gray value of 2 for the lesion, 1 for the one-pixel-wide boundary, and 0 for the rest of the image [5].

Figure 1.1 shows some examples of k-means segmented images obtained from the 888 dataset.

Some of the terms commonly used throughout this document are defined below for the convenience of the reader.

1. **Lesion:** The part of the image which represents the skin lesion, as opposed to background skin. This part has useful data that can be processed to obtain significant information. This is also usually the dark-colored area or region when compared to the skin.

2. **Lesion mask/border:** The area of the lesion separated from the image and stored as a binary mask.

3. **Common features:** The features are extracted either for the lesion or the segment in the lesion. Common features refer to those which are obtained from the lesion as a whole, i.e., which are common for all the segments.

4. **Segment features:** Features computed separately for the different segments or, in other words, features that are specific to the segment.

5. **Blob**: A group of connected segmented pixels based on eight-connectivity. Some blobs have a total area less than 25 pixels. These blobs are rarely significant and hence are discarded in some of the computations.

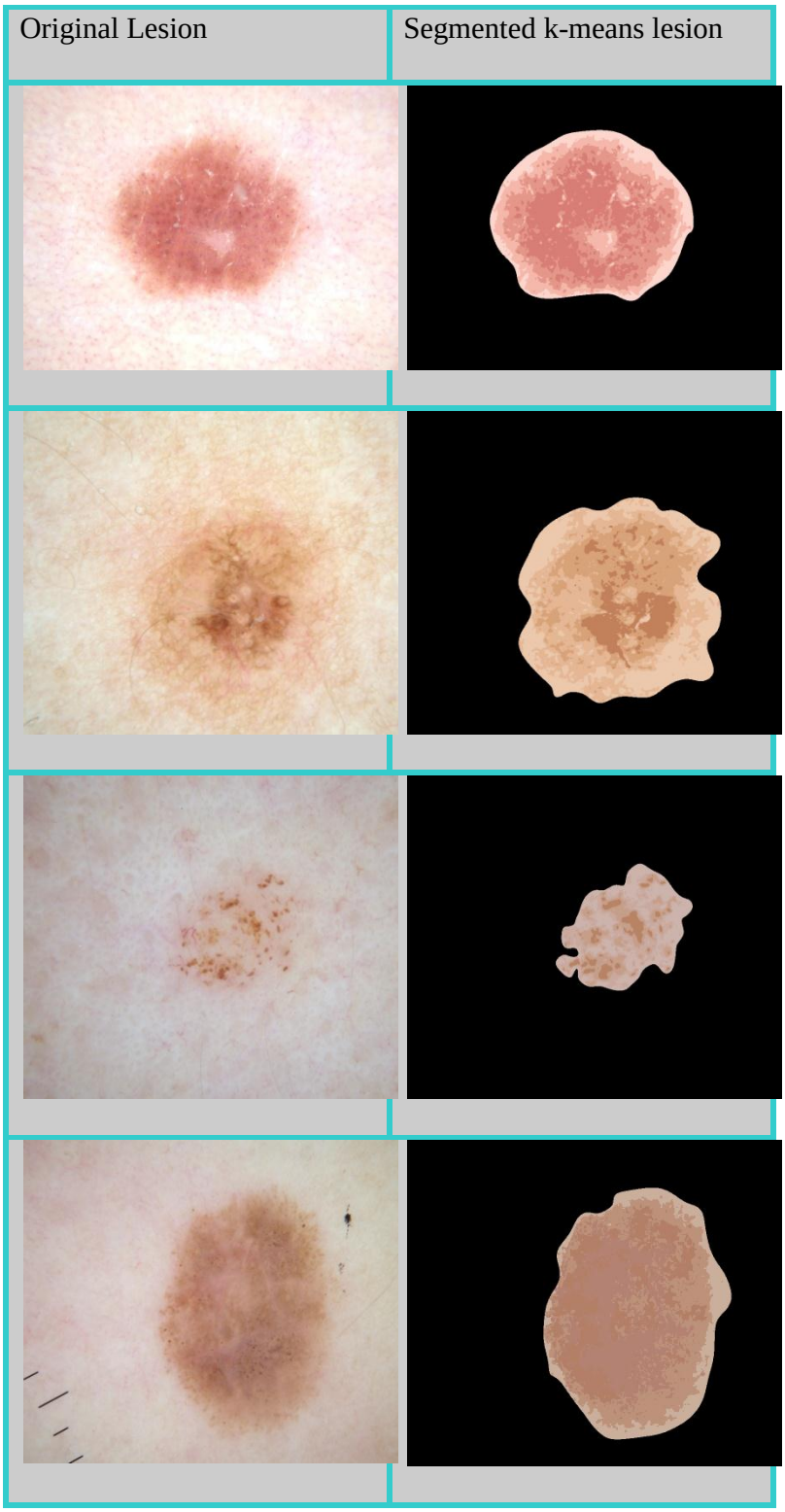

Figure 1.1: Original lesion images with k-means segmented images

2. FEATURE EXTRACTION

2.1 DATA EXTRACTION

The lesion area in each image was segmented into four different segments based on k-means clustering of color. The segment that had the darkest color was considered the first segment followed by the rest. The lighter segment as such is taken as the last segment and all four segments were used for feature extraction. Two-hundred-twenty-six features were obtained taking color as the basis, where six were the common features to all segments and the rest were found with respect to each different segment. As such 55 features were extracted from each of the four different segments which are described in the tables along with the common features. The four segments found in a sample image by k-means clustering are shown in Figure 2.1.

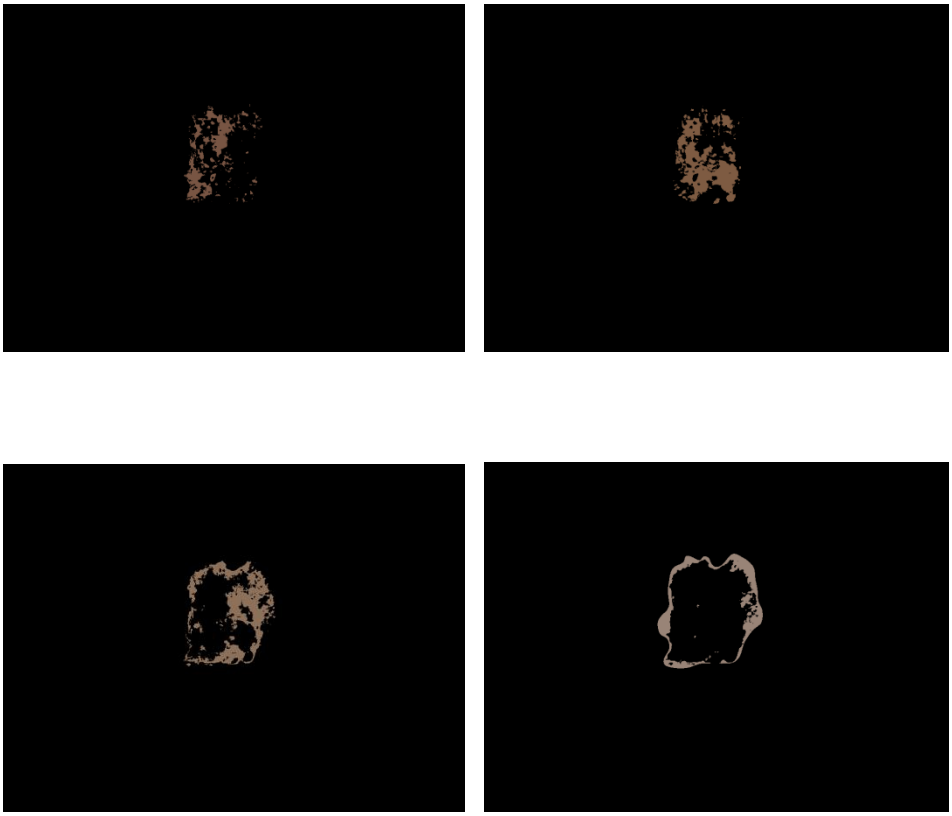

Figure 2.1: Image segments from k-means segmentation on a sample image

2.2 COMMON FEATURES

Table 2.1 lists the features extracted from the lesion as a whole. There are a total of six such features found for every lesion of the 888 data set. Descriptions of these features are explained in detail in the next section.

Table 2.1: Features extracted from the lesion area of the image

COMMON FEATURES
Ring value
Maximum segment in the peripheral ring
Last color in the peripheral ring
Total area of the lesion
X-component of the lesion centroid
Y-component of the lesion centroid

2.3 SEGMENT FEATURES

Apart from the common features listed in Table 2.1, every segment has its own set of values for features dependent on color. In other words, every segment has some specific features whose values vary depending on the segment under consideration. All the segment-specific features are listed in Table 2.2 with detailed explanations of each feature given in the next section.

Table 2.2: Segment-specific features found for each different segment

Percentage of segment in peripheral ring
Average red value of the segment before filling
Average green value of the segment before filling
Average blue value of the segment before filling
Number of blobs in the segment including the blobs having an area less than 25
Number of blobs in the segment excluding the blobs having an area less than 25
Area of the segment before filling
Area of the segment after filling
Area of the largest blob in the segment before filling
Area of the largest blob in the segment after filling
Perimeter of the segment before filling
Area of the segment after filling
Area of the largest blob in the segment before filling
Area of the largest blob in the segment after filling
Perimeter of the segment before filling
Perimeter of the segment after filling
Perimeter of the largest blob in the segment before filling
Perimeter of the largest blob in the segment after filling
Normalized perimeter of the segment before filling
Normalized perimeter of the segment after filling

Table 2.2: Segment-specific features found for each different segment (cont.)

Normalized perimeter of the largest blob in the segment before filling
Normalized perimeter of the largest blob in the segment after filling
Intensity drop for the internal perimeter of the segment before filling
Intensity drop for the internal perimeter of the segment after filling
Intensity drop for the internal perimeter for the largest blob in the segment before filling
Intensity drop for the internal perimeter for the largest blob in the segment after filling
Normalized intensity drop for the internal perimeter of the segment before filling
Normalized intensity drop for the internal perimeter of the segment after filling
Normalized intensity drop for the internal perimeter with the largest blob in the segment before filling
Normalized intensity drop for the internal perimeter with the largest blob in the segment after filling
External perimeter of the segment before filling.
External perimeter of the segment after filling.
External perimeter of the largest blob for the segment before filling.
External perimeter of the largest blob for the segment after filling.
Normalized external perimeter of the segment before filling.
Normalized external perimeter of the segment after filling.
Normalized external perimeter of the largest blob in the segment before filling.
Normalized external perimeter of the largest blob in the segment after filling
Intensity drop for external perimeter of the segment before filling.
Intensity drop for external perimeter of the segment after filling.
Intensity external parameter of the largest blob in the segment before filling.

Table 2.2: Segment-specific features found for each different segment (cont.)

Intensity external parameter of the largest blob in the segment after filling.
Normalized intensity drop external parameter for the segment before filling.
Normalized intensity drop external parameter for the segment after filling.
Normalized intensity drop external parameter for the largest blob in the segment before filling.
Normalized intensity drop external parameter for the largest blob in the segment after filling.
Centroid of the segment (x,y)
Eccentricity of the segment
Normalized distance for the segment
Luminance (Absolute and Background)
Average skin color (R,G,B) of the image
Average distance of the pixel
Average Red chromaticity of the segment
Variance of the red chromaticity of the segment
Standard deviation of the red chromaticity of the segment

The segments are formed on the basis of average R, G, and B values for pixels in the segment. However since the clustering was done in the color space, there were some pixels or blobs in the spatial domain that are present outside the primary area(s) of the segment under consideration.

For every segment there are blobs or objects of small area that can be discarded for a portion of the feature extraction process. As such, an area of 25 pixels has been set as the lower limit for the blob to be considered in some portions of the feature extraction process. Also some of the blobs were spread over a wide area with holes in them (where the holes are due to blobs from other segments). For this purpose, the blobs having an area greater than 25 and belonging to a segment have been filled to remove the small (area < 25 pixels) blobs from another segment. Both the calculations before filling and after filling are taken into account for some features.

3. FEATURES

The features listed Section 2 are described in detail below.

3.1 COMMON FEATURES

Table 3.1 shows the detailed explanation of all the common features for the lesion. The term "lesion border" means the one-pixel-wide border of the lesion (manually marked as described earlier).

Table 3.1: Description of features common to the lesion

1-Ring Value: $\frac{\text{common area of the segment present within the border of the lesion}}{\text{total number of pixels in the border of the lesion}}$ Intersection of the segment with the one pixel wide lesion border/peripheral ring divided by the total number of pixels in the one pixel wide border of the lesion. It can also be called the ratio of the segment present in the border. This value is computed for the four segments and the highest ring value is returned.
2-Maximum segment in the peripheral ring: The segment with the highest ring value is considered as the maximum segment. The segment number (value 1, 2, 3 or 4) is returned to indicate the maximum segment.
3-Last color in the peripheral ring: This feature returns a one if the segment 4 has the highest ring value for the given image. If the ring value is higher for any of the other three segments, 1, 2 or 3, this feature will return a value of zero.
4-Total area: Total number of the pixels in the lesion.

Table 3.1: Description of features common to the lesion (cont.)

5-Centroid (x-axis): Center coordinate (x- axis) for the whole lesion.
6-Centroid (y-axis): Center coordinate (y- axis) for the whole lesion.

3.2 SEGMENT FEATURES

Table 3.2 provides a detailed description of each of the features found separately with respect to each of the four different segments of the lesion. The first segment is the darkest which is determined by the intensity and similarly the other segments are determined. As the segments progress outwards the darkness decreases and the last one is the outermost segment or the light colored segment. In the features listed below, the ones that are found after filling do not consider smaller blobs with area less than 25 pixels in the computation. Those features that are termed as before filling take into account the smaller blobs which have an area less than 25 pixels. A new parameter has been introduced in this study to determine the color intensity levels difference between the perimeter and outside the perimeter for the segment image (one pixel wide). The MATLAB function “region props” is used to measure a set of properties for each connected component (object or blob) in the binary image of that segment. These properties can be shape or the pixel value measurements. This research work uses the region props for the measurement of area of the blobs, which falls under the shape measurements.

Also the first segment has no color in it and hence forth the external and internal new perimeters are the same values for first segment but different for other segments. The features

found including smaller blobs (without filling holes) are listed in the Table 3.2.

Table 3.2: Description of features computed including the smaller blobs

1-Percentage of segment in peripheral ring: This is the number of pixels of the respective segment (first/second/third/fourth) present in the one-pixel-wide lesion border.
2-Average red value for the segment: This returns the average red value for the respective segment.
3-Average green value for the segment: This returns the average green value for the respective segment.
4-Average blue value for the segment: This returns the average blue value for the respective segment.
5-Number of blobs in the segment before filling: This measures the total number of blobs present within the segment without holes being filled. These are determined using the area properties of the region props in scalar values in MATLAB.
6-Area of the segment before filling: Area of the segment in pixels with all the smaller blobs included and without filling the holes in the segment.
7-Area of the largest blob in the segment before filling: This parameter returns the area in pixels for the largest blob in the segment before the holes are filled.
8-Perimeter of the segment before filling: The internal perimeter for all the blobs in the segment before filling the holes. This includes the blobs having an area less than 25 and is found using the number of non-zero matrix elements in the segment.
9-Perimeter of the largest blob in the segment before filling: Internal perimeter for the largest blob present in the segment before the holes are filled.
10-Normalized perimeter of the segment before filling: Internal perimeter of the segment before the holes are filled divided by the square root of the lesion area.
11-Normalized perimeter of the largest blob in the segment before filling: Internal perimeter of the largest blob in the segment before the holes are filled divided by the square root of the lesion area.
12-Intensity drop for the internal perimeter of the segment before filling: This is the difference in the average color intensity levels between the pixels present along the border of the segment and those that are present one pixel outside the border. This parameter is found including the blobs having an area less than 25 and without filling the holes.

Table 3.2: Description of features computed including the smaller blobs (cont.)

13-Intensity drop for the internal perimeter for the largest blob in the segment before filling: This is the difference in the average color intensity levels between the pixels of the largest blob present along the border of the blob and in the largest blob present just one pixel wide outside the border. This parameter is found including the blobs having an area less than 25 and without filling the holes.
14-Normalized intensity drop for the internal perimeter of the segment before filling: New color intensity drop for the internal perimeter of the segment before filling divided by the square root of total area of the lesion.
15-Normalized intensity drop for the internal perimeter of the largest blob in the segment before filling: New color intensity drop for the internal perimeter of the largest blob in the segment before filling divided by the square root of total area of the lesion.
16-External perimeter of the segment before filling: Blobs belonging to one segment are also present lying outside the primary area(s) of the segment and their significance level is evaluated using this feature. The combined perimeter for all the areas of the segment under consideration is found. Then the number of non-zeros found from the external perimeter. This parameter is computed for the blobs having unfilled holes and also those that are having an area less than 25.
17-External perimeter of the largest blob in the segment before filling: External perimeter is determined for the largest blobs. The holes are unfilled here.
18-Normalized external perimeter of the segment before filling: External perimeter of the segment before filling divided by the square root of the total area of lesion.
19-Normalized external perimeter of the largest blob in the segment before filling: External perimeter of the largest blob in the segment before filling divided by the square root of the total area of lesion.
20-Intensity drop for external perimeter of the segment before filling: The difference in average color intensities between the blob present along the border of the external perimeter and the largest blob present outside by just one pixel wide. This feature is computed before the holes are filled for the segment.
21-Intensity drop for external perimeter of the largest blob in the segment before filling: This feature determines the average color difference for the largest blob along the border of the external perimeter and the one that is lying just one pixel outside the border, when the holes are not filled.

Table 3.2: Description of features computed including the smaller blobs (cont.)

22-Normalized intensity drop for external perimeter for the segment before filling: New color intensity drop for external perimeter of the segment before filling divided by the square root of the total area of the lesion.
23-Normalized intensity drop for external perimeter for the largest blob of the segment before filling: New color intensity drop for external perimeter of the largest blob in the segment before filling divided by the square root of the total area of the lesion.
24-Centroid of the segment (x axis): The x-coordinate of the segment is calculated using the region props properties.
25-Centroid of the segment (y axis): The y-coordinate of the segment is calculated using the region props properties.
26-Eccentricity of the segment: This is the distance between the centroid of each segment and the centroid of the lesion.
27-Normalized distance for the segment: This is the ratio of the distance between the centroid of each segment and the centroid of the lesion to the square root of the total area of the lesion.
28-Luminance (Absolute): A measure of the absolute brightness for the desired segment. Luminance = 0.30R+0.59G+0.11B
29-Luminance (Background): A measure of the background luminance for the desired segment. Luminance = 0.30R+0.59G+0.11B
30-Average skin color (R) of the image: This feature deals with finding the average red value from the skin which surrounds the lesion in the original lesion image. The surrounding skin region was extracted by overlaying the mask image on the non-skin image. This was then used to find the mean from the red region of the skin image.
31-Average skin color (G) of the image: This feature deals with finding the average green value from the skin which surrounds the lesion in the original lesion image.
32-Average skin color (B) of the image: This feature deals with finding the average blue value from the skin which surrounds the lesion in the original lesion image.
33--Average distance of the pixel: The average distance from each pixel in the segment to the centroid of the segment is found.

Table 3.2: Description of features computed including the smaller blobs (cont.)

34-Red chromaticity of the segment: This feature determines the average red chromaticity of the given segment. Red chromaticity is defined as $R_{ch} = \frac{R}{R+G+B}$.
35-Variance of the red chromaticity of the segment: This finds the variance of the red chromaticity of the segment.
36-Standard deviation of the red chromaticity of the segment

Table 3.3 continues the segment specific features for those features which do not include smaller blobs of area less than 25.

Table 3.3: Description of features computed excluding the smaller blobs

37-Number of blobs in the segment after filling: This feature returns the number of blobs present within the segment once the holes are filled. This also is found using region props.
38-Area of the segment after filling: Area of the segment in pixels after filling all the holes and discarding the smaller blobs.
39-Area of the largest blob in the segment after filling: Once the segment is filled the area of the largest blob in pixels is determined.
40-Perimeter of the segment after filling: After the smaller blobs are discarded and the holes are filled the internal perimeter of the segment is determined by finding the number of non-zeros matrix elements in the segment.
41-Perimeter of the largest blob in the segment after filling: Holes are filled and the internal perimeter of the largest blob is then calculated.
42-Normalized perimeter of the segment after filling: Internal perimeter of the segment after the holes are filled divided by the square root of the lesion area.
43-Normalized perimeter of the largest blob in the segment after filling: Internal perimeter of the largest blob in the segment after the holes are filled divided by the square root of the lesion area.
44-Intensity drop for the internal perimeter of the segment after filling: This is the difference in the average color intensity levels between the pixels present along the border of the segment and those that are present one pixel outside the border. This parameter is found excluding the blobs having an area less than 25 and after filling the holes.
45-Intensity drop for the internal perimeter of the largest blob in the segment after filling: Average color difference between the largest blobs lying along the border of the segment to the largest blob present just one pixel outside the segment. This contains only blobs having an area greater than 25 and those that have the holes filled.
46-Normalized intensity drop for the internal perimeter of the segment after filling: New color intensity drop for the internal perimeter of the segment after filling divided by the square root of the total area of the lesion.
47-Normalized intensity drop for the internal perimeter of the largest blob in the segment after filling: New color intensity drop for the internal perimeter of the largest blob in the segment after filling divided by the square root of the total area of the lesion.

Table 3.3: Description of features computed excluding the smaller blobs (cont.)

48- External perimeter for the segment after filling: This feature is found for the external perimeter with the holes filtered.
49-External perimeter of the largest blob for the segment after filling: Same as the above feature except that now it is only found for the largest blobs with the holes filled.
50-Normalized external perimeter of the segment after filling: External perimeter of the segment after filling divided by the square root of the total area of the lesion.
51-Normalized external perimeter of the largest blob in the segment after filling: External perimeter of the largest blob in the segment after filling divided by the square root of the total area of the lesion.
52-Intensity drop for the external perimeter of the segment after filling : This is the average color intensity difference along the border of the external perimeter for the segment and one pixel outside the segment.
53-Intensity drop for the external perimeter of the largest blob for the segment after filling: This features is calculated after the holes are filled and for the largest blob for the segment
54-Normalized intensity drop for the external perimeter of the segment after filling: New color intensity drop for the external perimeter of the segment after filling divided by the square root of the total area of the lesion.
55-Normalized intensity drop of the external perimeter of the largest blob of the segment after filling: New color intensity drop for the external perimeter of the largest blob in the segment after filling divided by the square root of the total area of the lesion.

4. RESULTS

As defined in Section 3, a total of 226 features are extracted from each lesion image. These features will next be used to classify a lesion as melanoma or benign. As such, data analysis is very important to determine those features that are significant in differentiating between the two classes, and for this purpose forward stepwise logistic regression as implemented in SAS was utilized in the research work.

4.1 FEATURE SELECTION

SAS (Statistical Analysis System) is software provided by the SAS Institute which enables one to perform data warehousing, statistical analysis and data mining, applications development among many other operations [9]. This software has many built-in procedures which come in handy for the statistical analysis of data especially with large amounts of data. Forward stepwise logistic regression is used for feature selection and model building for this work [10]. The SAS LOGISTIC procedure is used in this research work to implement this in order to determine the significance of all the features computed from the k-means segmented images.

The SAS analysis generates a list file which contains all features for the model under consideration by SAS [9]. SLENTRY and SLSTAY are two variables that determine what significance level will be included in the model. SLENTRY stands for the minimum value required for a feature to be included in the model and SLSTAY refers to the consistent value that the feature should have to stay in final model. For current study SAS analysis was done for different values of SLENTRY=SLSTAY=0.05, 0.1, 0.2, 0.35 and 0.5. Of all those, the results from 0.5 produce the model with the best diagnostic

accuracy at the expense of having a more complex model (a model based on a significantly larger number of features). However for comparison purposes, results from SLENTRY=SLSTAY=0.1 are also explained in addition to those of SLENTRY=SLSTAY=0.5 in the subsequent sections. The features that significantly help in the differentiation process are termed as significant features which are used for a later part of the data analysis. A table containing the estimate of correct melanomas and non-melanomas is also included in the SAS list file. With SLENTRY=SLSTAY=0.5 stepwise logistic regression returned 90 significant features whereas for SLENTRY=SLSTAY=0.1, 21 features were returned as significant. These significant features include some of the common features to the lesion and some of the features for the segments. These are listed in the respective sections below.

4.1.1 Significant Features Common to Lesion. Tables 4.1 & 4.2 show the common features found significant in the classification procedure for SLENTRY=SLSTAY=0.5 and 0.1 respectively.

Table 4.1: Significant features for the lesion with SLENTRY=SLSTAY=0.5

Ring Value
Maximum segment in the peripheral ring
Last color in the peripheral ring

Table 4.2: Significant features for the lesion with SLENTRY=SLSTAY=0.1

Total area
Centroid (x-axis)

4.1.2 Significant Features Specific to the Segment. Though the same features are extracted for all four segments, their values returned determine the significance of those features for the segment and the lesion as whole. As such the four different segments are listed with their significant features, starting with the segment having the darkest color in the lesion. The significant features for the first, second, third and fourth segments respectively are as listed below in Tables 4.3 to 4.10.

Table 4.3: Significant features for the first segment in the lesion with SLENTRY=SLSTAY=0.5

Percentage of segment in the peripheral ring
Number of blobs in the segment before filling
Area of the segment before filling
Area of the segment after filling

Table 4.3: Significant features for the first segment in the lesion with SLENTRY=SLSTAY=0.5 (cont.)

Area of the largest blob in the segment before filling
Area of the largest blob in the segment after filling
Normalized perimeter of the segment before filling
Normalized perimeter of the largest blob in the segment after filling
Intensity drop for the internal perimeter of the segment before filling
Intensity drop for the internal perimeter of the segment after filling
Intensity drop for the internal perimeter for the largest blob in the segment after filling
Normalized intensity drop for the internal perimeter of the segment before filling
Normalized intensity drop for the internal perimeter of the segment after filling
Normalized intensity drop for the internal perimeter with the largest blob in the segment after filling
Average skin color (R) of the image
Average skin color (G) of the image
Average distance of the pixel
Average Red chromaticity of the segment
Variance of the red chromaticity of the segment
Standard deviation of the red chromaticity of the segment

Table 4.4: Significant features for the first segment in the lesion with SLENTRY=SLSTAY=0.1

Average blue value for the segment
Area of the segment before filling
Variance of the red chromaticity of the segment
Standard deviation of the red chromaticity of the segment

Table 4.5: Significant features for the second segment in the lesion with SLENTRY=SLSTAY=0.5

Average red value of the segment before filling
Number of blobs in the segment including the blobs having an area less than 25
Number of blobs in the segment excluding the blobs having an area less than 25
Perimeter of the largest blob in the segment before filling
Normalized perimeter of the segment before filling
Normalized perimeter of the largest blob in the segment before filling
Intensity drop for the internal perimeter of the segment before filling
Intensity drop for the internal perimeter of the segment after filling
Intensity drop for the internal perimeter for the largest blob in the segment before filling
Normalized intensity drop for the internal perimeter of the segment before filling
External perimeter of the segment after filling.

Table 4.5: Significant features for the second segment in the lesion with SLENTRY=SLSTAY=0.5 (cont.)

External perimeter of the largest blob for the segment after filling.
Normalized external perimeter of the segment before filling.
Normalized external perimeter of the segment after filling
Normalized external perimeter of the largest blob in the segment before filling.
Normalized external perimeter of the largest blob in the segment after filling
Intensity drop for external perimeter of the segment before filling.
Intensity external parameter of the largest blob in the segment after filling.
Normalized intensity drop external parameter for the segment after filling.
Eccentricity of the segment
Average distance of the pixel
Average Red chromaticity of the segment

Table 4.6: Significant features for the second segment in the lesion with SLENTRY=SLSTAY=0.1

Percentage of segment in the peripheral ring
Average red value for the segment
Eccentricity of the segment

Table 4.7: Significant features for the third segment in the lesion with SLENTRY=SLSTAY=0.5

Percentage of segment in the peripheral ring
Number of blobs in the segment before filling
Area of the segment before filling
Area of the largest blob in the segment before filling
Perimeter of the segment before filling
Perimeter of the segment after filling
Perimeter of the largest blob in the segment before filling
Perimeter of the largest blob in the segment after filling
Normalized perimeter of the largest blob in the segment after filling
Intensity drop for the internal perimeter of the segment before filling
Intensity drop for the internal perimeter for the largest blob in the segment before filling
External perimeter of the segment before filling.
External perimeter of the largest blob for the segment before filling.
External perimeter of the largest blob for the segment after filling.
Intensity drop for external perimeter of the segment after filling.
Intensity drop for external parameter of the largest blob in the segment after filling.
Eccentricity of the segment
Normalized distance for the segment
Luminance (Absolute)
Standard deviation of the red chromaticity of the segment

Table 4.8: Significant features for the third segment in the lesion with SLENTRY=SLSTAY=0.1

Percentage of segment in the peripheral ring
External perimeter of the segment after filling.
Standard deviation of the red chromaticity of the segment

Table 4.9: Significant features for the fourth segment in the lesion with SLENTRY=SLSTAY=0.5

Percentage of segment in the peripheral ring
Average green value for the segment
Number of blobs in the segment before filling
Number of blobs in the segment after filling
Area of the segment after filling
Area of the largest blob in the segment after filling
Perimeter of the segment before filling
Normalized perimeter of the largest blob in the segment before filling
Intensity drop for the internal perimeter for the largest blob in the segment before filling
Intensity drop for the internal perimeter for the largest blob in the segment after filling
Normalized intensity drop for the internal perimeter of the segment before filling
Normalized intensity drop for the internal perimeter of the segment after filling
External perimeter of the segment before filling.
External perimeter of the segment after filling.

Table 4.9: Significant features for the fourth segment in the lesion with SLENTRY=SLSTAY=0.5 (cont.)

Normalized external perimeter of the segment before filling
Normalized external perimeter of the segment after filling
Normalized external perimeter of the largest blob in the segment before filling
Normalized external perimeter of the largest blob in the segment after filling
Intensity drop for external perimeter of the segment before filling.
Intensity drop for external parameter of the largest blob in the segment before filling.
Normalized intensity drop for external perimeter for the segment before filling
Centroid of the segment (x axis)
Centroid of the segment (y axis)
Average distance of the pixel
Average Red chromaticity of the segment

Table 4.10: Significant features for the fourth segment in the lesion with SLENTRY=SLSTAY=0.1

Average green value for the segment
Normalized perimeter of the segment after filling
Normalized perimeter of the largest blob in the segment after filling
External perimeter of the segment before filling.
External perimeter of the segment after filling.

Table 4.10: Significant features for the fourth segment in the lesion with SLENTRY=SLSTAY=0.1(cont.)

Normalized intensity drop for external parameter for the segment before filling.
Centroid of the segment (y axis)
Eccentricity of the segment
Average distance of the pixel

4.2 LOGISTIC REGRESSION

After collecting the significant features from the SAS analysis, the logistic regression model is used to verify if an image has been correctly identified as a melanoma or a benign lesion. Logistic regression or the logistic model is generally used for prediction of the probability of occurrence of an event by fitting data to a logistic function. The logistic function is defined as

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}, \text{where } z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k.$$

The values β and x are derived from the section Analysis of Maximum Likelihood estimates of the SAS output. β_0 is known as the intercept and $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$ are called the regression coefficients for $x_1, x_2, x_3, \cdots, x_k$, respectively. Each of these coefficients represents the size of the contribution of that feature. $f(z)$ represents the probability of outcome of any item and z represents the measure of the total contribution of all the independent variables used in the model [10].

Logistic regression is applied with the significant columns (features) based on their significance levels in the current model (estimate). Those lesions with *f(z)* < a threshold, for example $f(z) < 0.5$, are classified as benign and those with *f(z)* > the threshold are classified as melanoma. These automatic classifications are then compared with the clinical diagnosis in the case of the benign lesions and with the dermatopathologist's diagnosis in the case of the melanoma lesions.

4.3 CLASSIFICATION RESULTS

The sensitivity and specificity percentages depict the number of melanoma and benign lesions found from the model. Sensitivity can be defined as true positives divided by number of melanoma lesions, whereas specificity is defined as true negatives divided by number of benign lesions. The probability levels along with the specificity and sensitivity are as shown below in the Table 4.11.

Table 4.11: Specificity and Sensitivity for the probability levels with SLENTRY=SLSTAY=0.5

Probability Level	Specificity	Sensitivity
0	100	0
0.02	94.9	26.4
0.04	91.8	39.5
0.06	86.2	49.5

Table 4.11: Specificity and Sensitivity for the probability levels with SLENTRY=SLSTAY=0.5 (cont.)

Probability Level	Specificity	Sensitivity
0.08	82.6	58
0.1	80	63.8
0.12	76.4	68.5
0.14	75.4	72
0.16	72.3	75.5
0.18	71.8	77.6
0.2	70.3	79.2
0.22	69.7	81.2
0.24	67.2	82.3
0.26	64.6	83.4
0.28	61.5	84.3
0.3	57.4	85.4
0.32	57.4	86.1
0.34	55.9	86.9
0.36	55.9	87.2
0.38	54.9	87.7
0.4	52.8	88.7
0.42	51.3	89.2

Table 4.11: Specificity and Sensitivity for the probability levels with SLENTRY=SLSTAY=0.5 (cont.)

Probability Level	Specificity	Sensitivity
0.44	50.3	89.5
0.46	49.7	89.6
0.48	49.2	89.8
0.5	48.2	90.5
0.52	47.7	91.2
0.54	46.2	92.1
0.56	45.1	92.2
0.58	45.1	92.9
0.6	43.1	93.7
0.62	41.5	94.1
0.64	41	94.4
0.66	38.5	94.9
0.68	37.4	95.8
0.7	36.4	95.8
0.72	35.9	96.5
0.74	35.4	96.7
0.76	33.3	97
0.78	32.3	97.3
0.8	31.8	97.4
0.82	31.8	97.4

Table 4.11: Specificity and Sensitivity for the probability levels with SLENTRY=SLSTAY=0.5 (cont.)

0.84	29.7	97.5
0.86	28.7	98.4
0.88	27.7	99
0.9	26.7	99
0.92	23.6	99
0.94	21	99.4
0.96	17.9	99.6
0.98	12.3	99.7
1	0	100

4.4 RECEIVER OPERATING CHARACTERISTICS CURVE

The algorithm accuracy was tested by plotting the receiver operating characteristics (ROC) curve, which is a plot of sensitivity versus one minus specificity. The significant columns are different for different SLENTRY=SLSTAY values along with the sensitivity and specificity values. The area under the ROC curve is calculated to verify the accuracy of the SAS model for five different cases of SLENTRY=SLSTAY. The area under the ROC curve for SLENTRY=SLSTAY=0.05, 0.1, 0.2, 0.35 and 0.5 was calculated to be 0.821, 0.849, 0.856, 0.872 and 0.902 respectively.

Figures 4.1 and 4.2 show the ROC curves for SLENTRY=SLSTAY=0.5 and SLENTRY=SLSTAY=0.1, respectively, along with the area under the ROC curves (AUC).

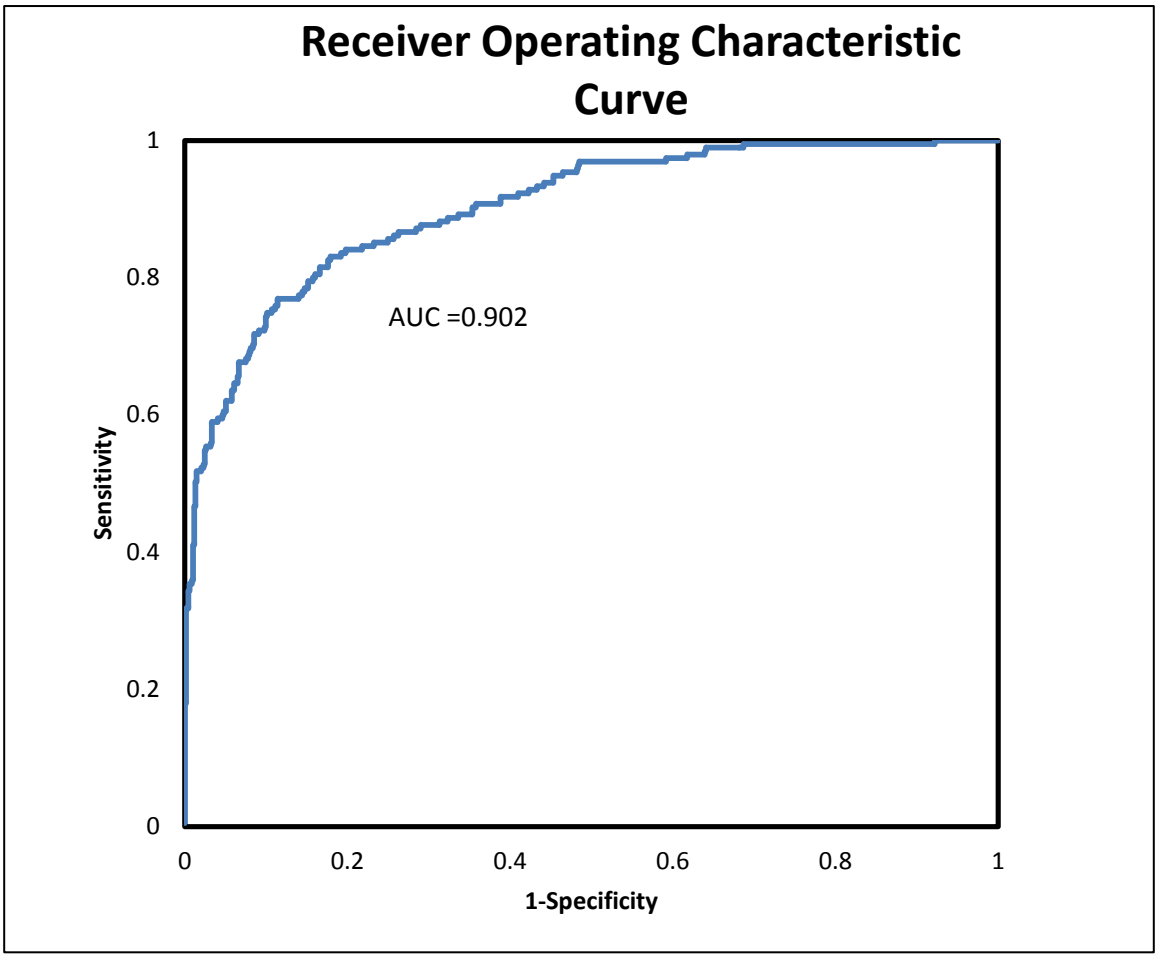

Figure 4.1: ROC curve for SLENTRY=SLSTAY =0.5

The ROC curve in Figure 4.1 was generated for SLENTRY=SLSTAY equal to 0.5. The area under the ROC curve (which is also given by the value of the SAS parameter c) is

0.902. The plot in Figure 4.2 shows the ROC curve for alpha = 0.1, with an area under the ROC curve equal to 0.849. When the alpha value is decreased from 0.5 to 0.1, the size of the model decreases from 90 features to 21 features and the diagnostic accuracy drops some as measured by the areas under the ROC curves.

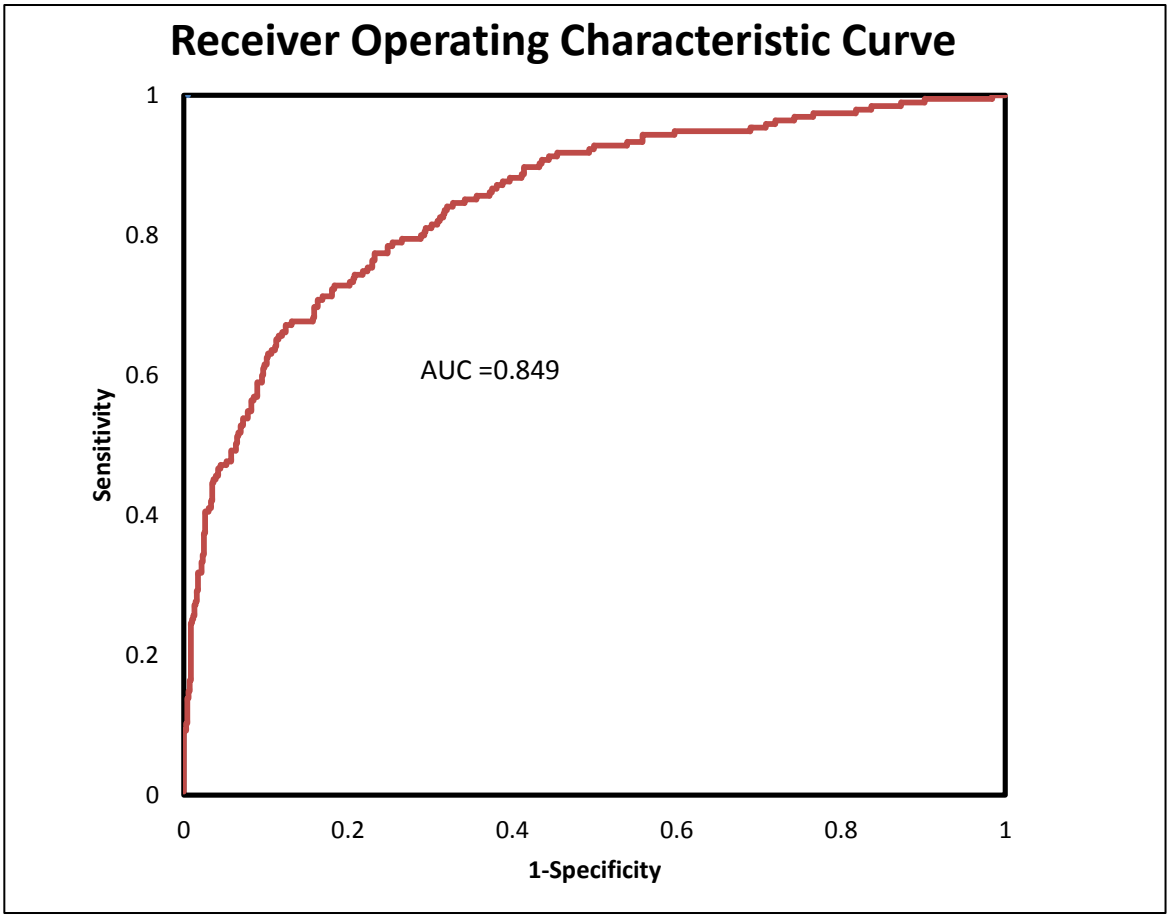

Figure 4.2: ROC curve for SLENTRY=SLSTAY =0.1

5. CONCLUSIONS

The current study examines the use of k-means clustering to segment dermoscopy images based on color to produce features that can help in distinguishing between melanoma and benign tumors. This was achieved using k-means segmentation in this work with the value of k equal to 4. The lesions are segmented using color as the basis and then these segments are used for the data extraction.

A data set of 888 dermoscopy images, containing 195 melanoma and 693 benign lesions, was used. This produced 226 features which were analyzed using forward stepwise logistic regression, yielding a set of 90 significant features. The model based on these features gave a parameter estimate of 90.2 % through SAS.

The area under the ROC curve was equal to the SAS estimate with a value of 0.902.

6. FUTURE WORK

It might be possible to find automatically an optimum or near-optimum value of k for the color segmentation of a given dermoscopy image based on statistical data obtained from the original image. Accuracy can be further enhanced using different color spaces for segmenting the skin lesions. This has the potential to help in improving the accuracy of the results. Also introducing several additional features may improve the classification results.

REFERENCES

[1] Rogers, H.W., Weinstock, M.A., Harris, A.R., et al. Incidence estimate of non-melanoma skin cancer in the United States, 2006. *Arch Dermatol* 2010; **146**(3):283-287.

[2] Siegel, R., Naishadham, D., and Jemal, A. (2012). Cancer Statistics, 2012. *CA: A Cancer Journal for Clinicians*, **62**(1), 10-29.

[3] Mocellin, S., and Nitti, D. (2011). Cutaneous melanoma in situ: translational evidence from a large population-based study. *Oncologist*, **16**(6), 896-903.

[4] Hamerly, G. and Elkan, C. (2002). "Alternatives to the k-means algorithm that find better clusterings". *Proceedings of the eleventh international conference on Information and knowledge management (CIKM)*,**H.3.3, I.5.3**,600-607

[5] Meka, L.S. (2012). White area analysis for detection of malignant melanoma, M.S. Thesis, Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO.

[6] Menzies, S.W. and Zalaudek, I. (2006). Why perform dermoscopy? The evidence for its role in the routine management of pigmented skin lesions. *Archives of Dermatology*, **142**, 1211-1222.

[7] Liang Wang, Christopher Leckie, Kotagiri Ramamohanarao, James Bezdek (2009), Automatically Determining the Number of Clusters in Unlabeled Data Sets, *IEEE Transactions on Knowledge and Data Engineering*, **21**,(3),335-350.

[8] Srinivasulu Aside ,Subba Rao Ch D.V, Saikrishna V. (2010). Finding the Number of Clusters in Unlabeled Datasets using Extended Dark Block Extraction, *International Journal of Computer Applications*, **7**(3), 0975 – 8887.

[9] Introduction to SAS. UCLA: Academic Technology Services, Statistical Consulting Group http://www.ats.ucla.edu/stat/sas/. Last referred 03/12/2012

[10] Agresti, A. (2007), Building and applying, logistic regression models, *An Introduction to Categorical Data Analysis*, Hoboken: Wiley, p. 138.

VITA

Snigdha Priya Bommadevara was born in Andhra Pradesh, India. In May 2011, she received her Bachelor`s degree in Electrical Engineering from the Joginapally BR Engineering College, Hyderabad, Andhra Pradesh, India. She worked as a research assistant with Dr. Randy Moss on the project – 'Feature Extraction through K-means Segmentation for Melanoma Detection'. She worked as a Graduate Teaching Assistant during her Master's in 2012, where she was an instructor for Discrete Linear Systems lab in Electrical Engineering department. She worked as a Wi-Fi Testing Intern in Broadcom, Sunnyvale, CA, USA. In December 2013, she received her M.S. degree in Electrical Engineering from the Missouri University of Science and Technology, Rolla, Missouri, USA.