Missouri S&T Library and Learning Resources

Scholars' Mine

Masters Theses

Student Theses and Dissertations

Spring 2017

Automatic discrimination of basal cell carcinoma from sebaceous hyperplasia based on image processing of dermoscopy images

Jonathan Liao

Follow this and additional works at: https://scholarsmine.mst.edu/masters_theses

Part of the Electrical and Computer Engineering Commons

Department:

Recommended Citation

Liao, Jonathan, "Automatic discrimination of basal cell carcinoma from sebaceous hyperplasia based on image processing of dermoscopy images" (2017). *Masters Theses*. 8053.
https://scholarsmine.mst.edu/masters_theses/8053

This thesis is brought to you by Scholars' Mine, a service of the Missouri S&T Library and Learning Resources. This work is protected by U. S. Copyright Law. Unauthorized use including reproduction for redistribution requires the permission of the copyright holder. For more information, please contact scholarsmine@mst.edu.

AUTOMATIC DISCRIMINATION OF BASAL CELL CARCINOMA FROM SEBACEOUS HYPERPLASIA BASED ON IMAGE PROCESSING OF DERMOSCOPY IMAGES

by

JONATHAN LIAO

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

2017

Approved by

Dr. Randy H. Moss, Advisor

Dr. William V. Stoecker

Dr. R. Joe Stanley

Dr. Bijaya Shrestha

© 2017

Jonathan Liao

All Rights Reserved

ABSTRACT

This project focuses on discriminating basal cell carcinoma from sebaceous hyperplasia using image processing techniques. Basal cell carcinoma is a kind of malignant skin cancer that needs to be treated; however, if diagnosed as sebaceous hyperplasia, a benign lesion which is a mimic of basal cell carcinoma, then it may not be treated properly. Through observation, white pouch-like areas within the lesion appear in sebaceous hyperplasia images; whereas in basal cell carcinoma images, white areas tend to be formed in a smashed irregular figure shape. Hence, utilizing image processing techniques to segment these white areas from the images and using the resulting blob mask images to extract features to train a model for classification is the aim of the project to achieve a higher chance of correctly classifying basal cell carcinoma from sebaceous hyperplasia automatically through dermoscopy images.

ACKNOWLEDGMENTS

I would like to thank Dr. Moss for giving me the great foundation for conducting research that is based on image processing and for caring for me as a student who has just started to explore the world of engineering. Opening the door towards the world of research and image analysis, Dr. Moss has given me the tools for conducting research for this project and every research in the future ahead. I would like to thank Dr. Stoecker for giving professional input in a dermatologic perspective and valuable advices throughout the project. Dr. Stoecker has given me irreplaceable on field experience. He provided me the opportunity to participate in sampling some of the data that was used in the project. I would also like to thank Dr. Stanley and Dr. Shrestha for giving valuable advice on improving the project and thesis. Their efforts and their different point of view has perfected the project even more and have brought the project onto another whole new level.

I would also want to thank Jason, Nabin, Pramada and Kanakam for being wonderful superiors in the lab. They were all so generous on giving me confidence and help in numerous ways. I would like to like to thank Jason for using so much effort in giving inputs and thought in the project, and for building me up as a solid programmer and an individual engineer. I would like to thank the people of The Dermatology Center in Rolla and the patients that helped the sampling process for the research. Their kindness and love encouraged me to do my best in the sampling process and during the research.

I would like to use this chance to thank God for sending all the wonderful people that supported me throughout the whole pursue of my master's degree. Thank you Lord for giving strength, peace, and joy to me when I faced high obstacles, stormy days, and new challenges. I would like to thank my family members, Simon, Linda, Joshua, Sarah, and Sandy for giving me love and support every day. Thank you all that was with me on this exciting journey and may God bless you all.

TABLE OF CONTENTS

Page

ABSTRACT iii

ACKNOWLEDGMENTS iv

LIST OF ILLUSTRATIONS vii

LIST OF TABLES viii

SECTION

1. INTRODUCTION 1

2. METHODS 3

2.1. METHODOLOGY OVERVIEW 3

2.2.DATASET AND LESION BORDERS 4

2.2.1. Obtaining Lesion Image Set 4

2.2.2. Generating Manual Lesion Borders 4

2.3.IMAGE PROCESSING AND SEGMENTATION 5

2.3.1. Principal Component Transform 5

2.3.2. Thresholding Using Otsu Threshold Technique 6

2.3.3. Remove Noise from Binary Images 7

2.3.4. Area Filtering 7

2.4.EXTRACTING FEATURES 8

2.4.1. Color Features 9

2.4.2. Shape Features 11

3. FEATURE SELECTION 14

3.1.SAS 14

3.2.MODEL GENERATION AND FEATURE SELECTION 14

4. RESULTS 15

4.1.MODEL GENERATION RESULTS 15

4.2.EVALUATIONS ON MODELS AND TEST RESULTS 17

5. CONCLUSION 28

5.1.RESULTS AND DISCUSSIONS 28

5.2.LIMITATIONS AND FUTURE WORK 28

BIBLIOGRAPHY 30

VITA 31

LIST OF ILLUSTRATIONS

Page
Figure 1.1 Flow Chart of this Project..... 2
Figure 2.1 (a) Original image (b) After applying manual lesion border..... 5
Figure 2.2 (a) Original lesion image (b) PCT image 6
Figure 2.3 (a) PCT image (b) Binary image after applying Otsu threshold 6
Figure 2.4 (a) PCT image (b) Binary image after applying resulting threshold..... 7
Figure 2.5 (a) Binary image (b) After applying Opening Morphology operation..... 7
Figure 2.6 (a) 0.15% area filtered blob mask (b) 0.10% area filtered blob mask
(c) 0.06% area filtered blob mask (d) 0.02% area filtered blob mask 8
Figure 4.1 Data Set 1 (Left) and Data Set 2 (Right) Training Model ROC Curve..... 19
Figure 4.2 Data Set 3 (Left) and Data Set 4 (Right) Training Model ROC Curve..... 19
Figure 4.3 Data Set 5 (Left) and Data Set 6 (Right) Training Model ROC Curve..... 19
Figure 4.4 Data Set 7 (Left) and Data Set 8 (Right) Training Model ROC Curve..... 19
Figure 4.5 Data Set 1 (Left) and Data Set 2 (Right) Test Set ROC Curve 20
Figure 4.6 Data Set 3 (Left) and Data Set 4 (Right) Test Set ROC Curve 20
Figure 4.7 Data Set 5 (Left) and Data Set 6 (Right) Test Set ROC Curve 20
Figure 4.8 Data Set 7 (Left) and Data Set 8 (Right) Test Set ROC Curve 20

LIST OF TABLES

Page

Table 2.1 Color Features 9
Table 2.2 Color Planes 10
Table 2.3 Color Feature Calculations 11
Table 2.4 Shape Features 12
Table 2.5 Shape Feature Calculations I 12
Table 2.6 Shape Feature Calculations II 13
Table 4.1 Data Sets Used 15
Table 4.2 Area Under Curve and Features Chosen 16
Table 4.3 Features Chosen 17
Table 4.4 Measurements and Formulas 18
Table 4.5 Thresholds Used for Classification 21
Table 4.6 Number of Misclassified Images for Threshold 1 21
Table 4.7 Number of Misclassified Images for Threshold 2 22
Table 4.8 Number of Misclassified Images for Threshold 3 22
Table 4.9 Number of Misclassified Images for Threshold 4 23
Table 4.10 Results of Measurements of Training Model for Threshold 1 23
Table 4.11 Results of Measurements of Training Model for Threshold 2 24
Table 4.12 Results of Measurements of Training Model for Threshold 3 24
Table 4.13 Results of Measurements of Training Model for Threshold 4 25
Table 4.14 Results of Measurements of Test Set Results for Threshold 1 25
Table 4.15 Results of Measurements of Test Set Results for Threshold 2 26
Table 4.16 Results of Measurements of Test Set Results for Threshold 3 26
Table 4.17 Results of Measurements of Test Set Results for Threshold 4 27

1. INTRODUCTION

Skin cancer is the most common of all types of cancer, and basal cell carcinoma (BCC) (along with squamous cell carcinoma) is the most common type of skin cancer. Basal cell carcinoma often can be found around head and neck areas, especially areas that have been over exposed to sunlight, x-ray or other types of radiation. People that have fair skin, red hair or blond hair and eyes that are blue, grey or green tend to have the highest chance of developing basal cell carcinoma [1]. Approximately 2.7 million cases of basal cell carcinoma are diagnosed each year in the United States. The cost of surgical treatment for BCC is increasing the burden on healthcare systems [2]. Because BCC often appears on the face, surgery entails significant pain and potential for deformity.

Sebaceous hyperplasia (SH), in contrast, is a benign skin condition that is relatively common among middle-aged or older adults. Sebaceous hyperplasia is most commonly found on the face [3]. Unfortunately, sebaceous hyperplasia can closely mimic BCC. If BCC is diagnosed as SH, it may not be treated in time; conversely, if SH is diagnosed as BCC, an unnecessary biopsy may be performed.

Therefore, this project aims to automatically discriminate BCC from SH based on image processing of dermoscopy images to assist general practitioners and physician extenders, and perhaps eventually patients with the proper equipment, in distinguishing the two types of skin growths. Through observations, white pouch-like areas within the lesion area can be found in dermoscopy images of SH, and white areas with smashed irregular shapes can be found in dermoscopy images of BCC. In the past [4], research has been done on the topic of automatic detection and classification of dermoscopic structures. In this project, the focus is to distinguish BCC from SH based on the shape of the white lesion areas and other information from the dermoscopy images. Hence, an image processing approach was used in this research to classify the two types of dermoscopy images with high accuracy.

In the project, the principal component transform (PCT) technique is utilized for increasing contrast between the white areas and surrounding lesion areas. The Otsu threshold technique [5] is utilized to obtain an optimal threshold for segmenting the white areas within the lesion area. The two procedures generate blob masks for the white areas

within the lesions. Then, an opening morphology is applied to remove small areas of noise within the blob masks. To further filter out relatively small noise areas and to select target white areas, various area filters are applied for comparison to gain an optimal solution. To maintain size invariance, the area filters are based on a parameter that represents a percentage of the lesion area to generate different sets of blob masks for feature extraction.

After applying the blob masks to the original lesion images, color features [6] and shape features [7] from lesion areas within the blob masks are extracted. These features are presented to a logistic regression classifier to develop and train a model to distinguish BCC images from SH images. Section 2 explains the methodology used, processing stages and feature extraction in this project; Section 3 shows the feature selection method used for training models; Section 4 displays the results of the project and Section 5 gives the conclusion for the project. Figure 1.1 shows the flow chart of this project.

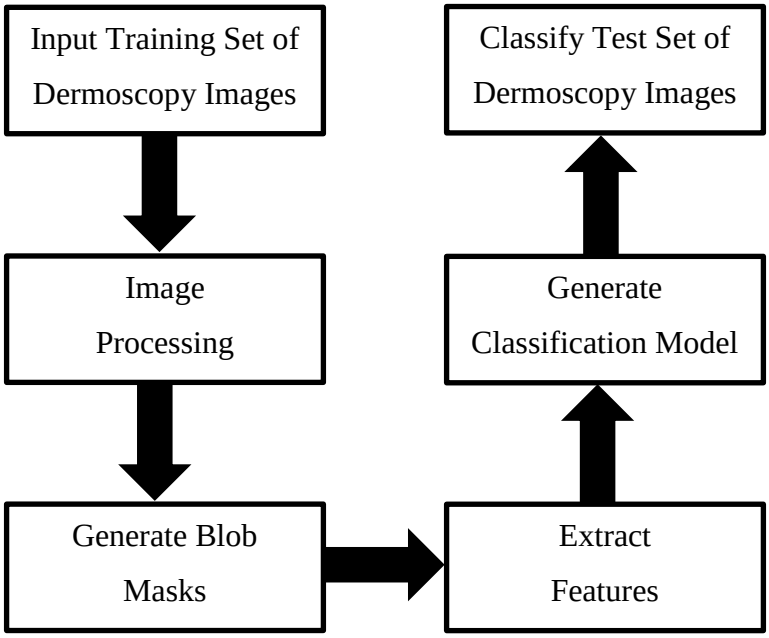


Figure 1.1 Flow Chart of this Project

2. METHODS

2.1. METHODOLOGY OVERVIEW

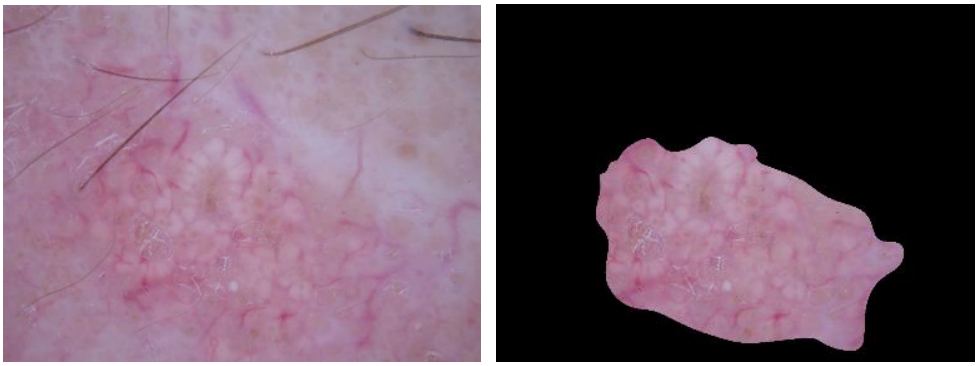
The following steps are the methods used in this project for discriminating BCC images from SH images:

1. Collect samples of BCC and SH dermoscopy images
2. Generate the manual lesion borders for all the dermoscopy images
3. Perform principal component transform on the bordered images
4. Generate blob masks using the Otsu threshold technique on the PCT images
5. Remove small noise objects by applying an opening morphology operation on blob masks
6. Apply various area filters to generate different sets of blob masks
7. Extract color and shape features from lesion areas within the blob masks
8. Choose samples from both sets of images to create a training set and test set for BCC and SH dermoscopy images
9. Use a logistic regression classifier to generate a model from the training set
10. Use the trained model on the test set to classify BCC and SH images
11. Calculate measurements based on the results to evaluate the classification ability of the trained model

2.2. DATASET AND LESION BORDERS

2.2.1. Obtaining Lesion Image Set. The datasets of BCC and SH that were used in this project were obtained by using a digital camera with a special dermoscopy lens attachment that focuses on the skin. The lens attachment is equipped with surrounding LED lights that illuminate the skin surface at a low angle of incidence, which when coupled with liquid or gel immersion enables the camera to detect subsurface structures. The resolution of the dermoscopy images is 1024 x 768; images that have larger resolutions are resized to 1024 x 768. The image set consists of 383 dermoscopy images; 348 of which are BCC images and 35 of which are SH images. The dermoscopy images were provided from four clinics as described in Kaur et al. [8]. The training set consists of 333 BCC images and 20 SH images, and the test set consists of 15 BCC images and 15 SH images. A random number generator was used to determine training and test sets from the complete data set for BCC; the SH training and test set is determined by the date of sampling, early samples went to the training set and new samples went to the test set. The training and test sets are disjoint.

2.2.2. Generating Manual Lesion Borders. To focus the search within the lesion area of the dermoscopy images, it is crucial to remove surrounding skin areas within the images that are not part of the lesion. To remove these areas from the image, lesion borders are to be determined and lesion border masks are generated. The lesion border masks are generated by using the Winshow program developed at the Missouri University of Science and Technology. The masks are manually determined by a user who selects points along the borders of the lesion in the image; the points are connected with a second-order spline; finally, the lesion border masks are then saved as images and checked by a dermatologist. In Figure 2.1, the original image and the image after applying the manual lesion border mask onto the original image are displayed. The image used to demonstrate the image processing stage is one of the SH images used in this project.



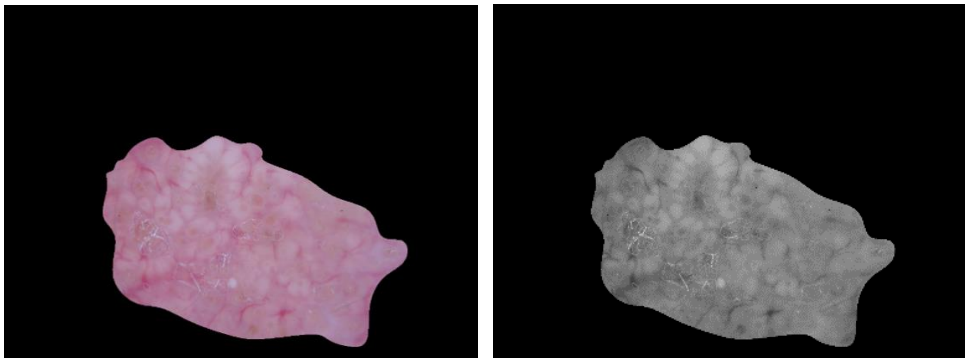
(a) (b)

Figure 2.1 (a) Original image (b) After applying manual lesion border

2.3. IMAGE PROCESSING AND SEGMENTATION

Recognizing the white areas directly from the original images is difficult due to color information within the lesion areas that suffer from low contrast. However, if the lesion area is preprocessed before the feature extraction process, the chances of accurately recognizing the white areas will increase.

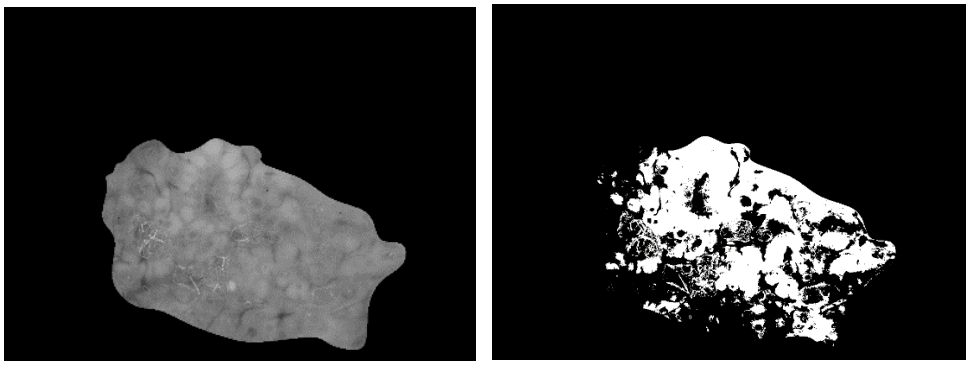
2.3.1. Principal Component Transform. In past research [9], the principal component transform has been used for image processing for assisting in the detection of target areas. The principal component transform used in this project finds a combination of the red pixel value, green pixel value and blue pixel value to form the highest contrast gray image possible. The eigenvector that corresponds to the largest eigenvalue in the principal component analysis (PCA) calculation can form a resulting image that has the greatest variance. Using the eigenvector values as weights for the red, green and blue pixel values in the original image, we form the gray lesion image with highest contrast. The goal in this process is to enhance the contrast between the white areas and its surrounding areas as shown in Figure 2.2.



(a) (b)

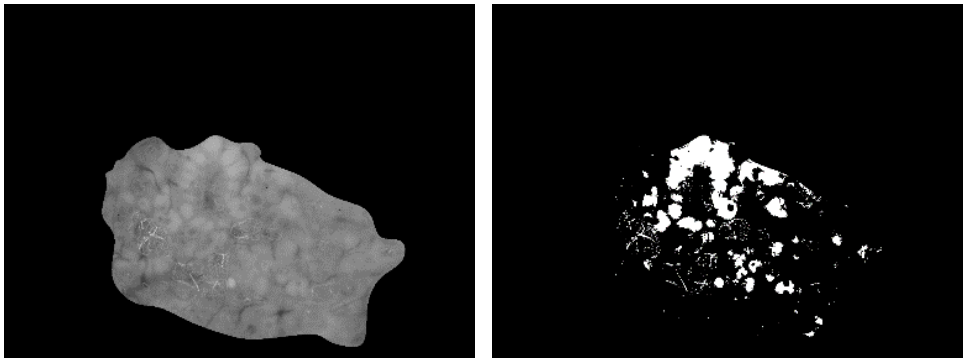
Figure 2.2 (a) Original lesion image (b) PCT image

2.3.2. Threshold Using Otsu Threshold Technique. After applying the principal component transform, the contrast between the white areas and the surrounding areas is increased. The next step is to determine the optimal gray value for thresholding the PCT images to obtain most of the white areas. The Otsu threshold [5] has been used extensively in image processing for finding the optimal gray value threshold in the image histogram. Applying the Otsu threshold directly to the PCT images is not effective in segmenting the white areas as shown in Figure 2.3. Hence, the Otsu threshold is used for an initial threshold value and upon that threshold value we add one standard deviation of the gray pixel values in the lesion area of the PCT images. The resulting threshold is effective in finding white areas in the PCT images. In Figure 2.4, the binary image is obtained by applying the resulting threshold to the lesion area within the PCT images.



(a) (b)

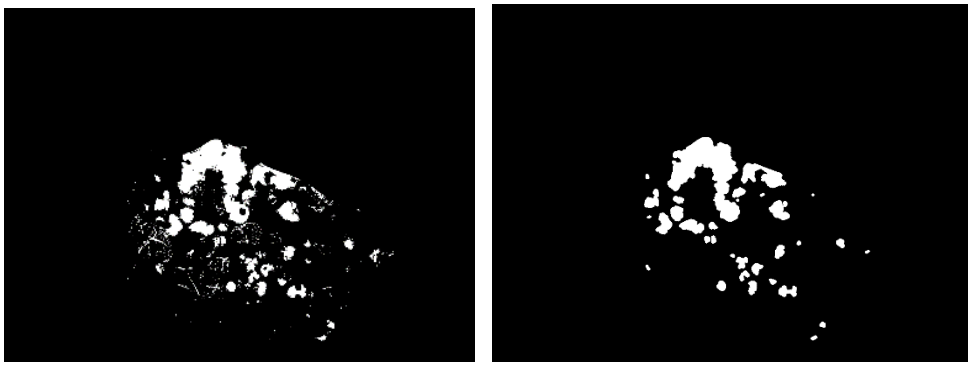
Figure 2.3 (a) PCT image (b) Binary image after applying Otsu threshold



(a) (b)

Figure 2.4 (a) PCT image (b) Binary image after applying resulting threshold

2.3.3. Removing Noise from Binary Images. The resulting binary images can form pouch-like areas for the sebaceous hyperplasia dermoscopy images; however, there are small noise areas that interfere with our focus on the pouch-like areas. To eliminate the small noise areas, a 7 by 7 circular structured kernel is applied for an opening morphology operation on the binary images. In the resulting images, most of the small noise areas are removed without eliminating any desired blobs, as shown in Figure 2.5.

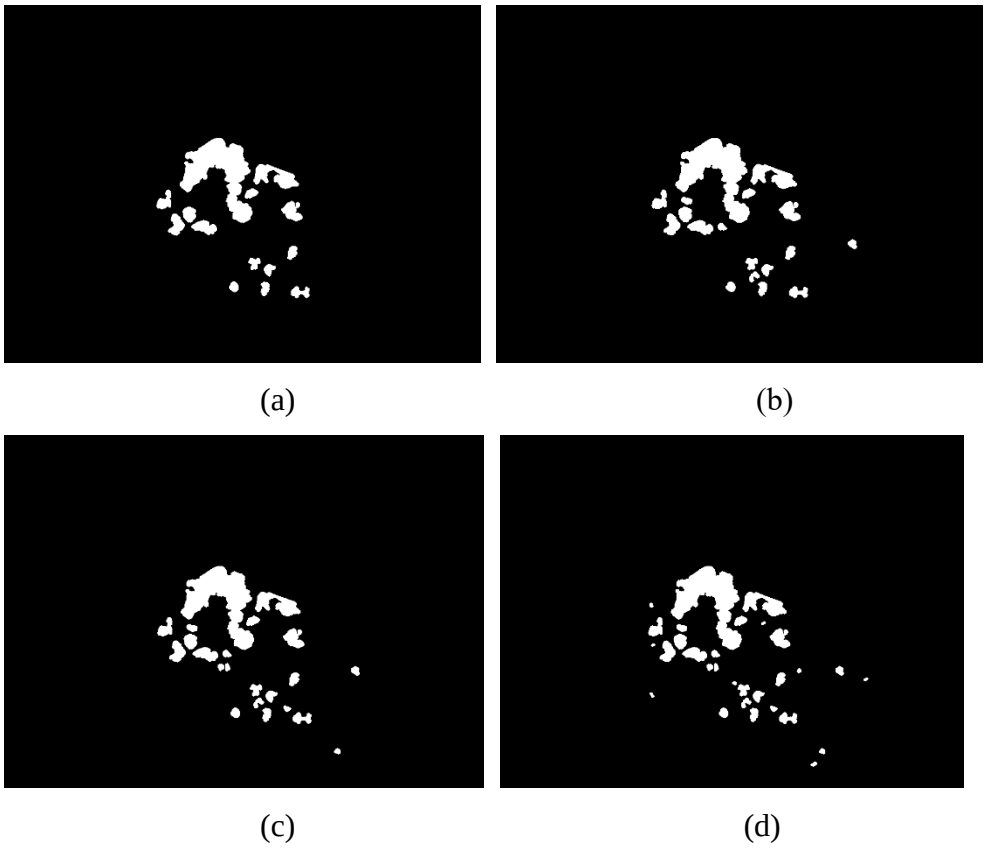


(a) (b)

Figure 2.5 (a) Binary image (b) After applying Opening Morphology operation

2.3.4. Area Filtering. As shown in Figure 2.5, there are still relatively small noise areas that can interfere with our focus on the pouch-like areas. To remove this noise adaptively, any blob that has area ≤ a threshold percentage of the total lesion area is removed. However, if the percentage of the area filter is too large, it will remove target white pouches. By

experimenting with different percentage values for the area filtering process, 0.15 percent, 0.1 percent, 0.06 percent, and 0.02 percent of the lesion area were selected for further study. Figure 2.6 shows the resulting images of the four area filters for the lesion shown in Figures 2.2–2.5. Through the area filter procedure above, four sets of blob masks were generated.



(a) (b)

(c) (d)

Figure 2.6 (a) 0.15% area filtered blob mask (b) 0.10% area filtered blob mask (c) 0.06% area filtered blob mask (d) 0.02% area filtered blob mask

2.4. EXTRACTING FEATURES

The features extracted are calculated and saved in a csv file. The features can be classified into color features and shape features. Color features are extracted from the overall lesion area and from the lesion areas within the blob mask. This generates two sets of 21 color features for each image, as shown in Table 2.1. The overall lesion area represents the whole lesion which is the lesion area within the manual lesion borders. Shape features are extracted only from the blob masks which generates one set of 9 shape features.

2.4.1. Color Features. Images contain color information useful for classifying and recognizing objects. In this project, three color spaces were used, the RGB color space, the HSV color space and the La*b* color space. Calculating the average and standard deviation for all nine color components results in 18 feature values which are saved in a csv file. Because the H variable "hue" is an angular value which has a discontinuity in the range encountered in skin images, the trigonometric functions sine and cosine of hue average and angular dispersion of hue average are also saved, giving a total of 21 color features. The formulas for obtaining the color planes are shown in Table 2.2 and the formulas for calculating the values of the color features are in Table 2.3. Since most of the images' hue value is close to red, which is the range that crosses the boundary of the value 0, there is a discontinuity in the hue variable. The feature "hue average" is obtained by rotating the hue average by 180 degrees for the model.

Table 2.1 Color Features

Color Space	Features
RGB	Red Average, Green Average, Blue Average, Red Standard Deviation, Green Standard Deviation, Blue Standard Deviation
HSV	Hue Average, Saturation Average, Value Average, Hue Standard Deviation, Saturation Standard Deviation, Value Standard Deviation, Angular Dispersion, Sine Value of Hue Average, Cosine Value of Hue Average
La*b*	L Average, a* Average, b* Average, L Standard Deviation, a* Standard Deviation, b* Standard Deviation

Table 2.2 Color Planes [6]

Color Planes	Color Plane Formula
RGB	Red plane, Green plane and Blue plane information obtained through the three corresponding channels of the image
HSV	$\mathbf{V = max\ (R, G, B)}$ $\mathbf{S} = \begin{cases} \frac{V - min(R,G,B)}{V}, & \text{if V} \neq 0 \\ 0, & \text{if V} = 0 \end{cases}$ $\mathbf{H} = \begin{cases} \frac{60(G-B)}{(V-min(R,B,G))}, & \text{if V} = \text{R} \\ \frac{120+60(B-R)}{(V-min(R,B,G))}, & \text{if V} = \text{G} \\ \frac{240+60(R-G)}{(V-min(R,B,G))}, & \text{if V} = \text{B} \end{cases}$
La*b*	$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$ **X = X / 0.950456** **Z = Z / 1.088754** $\mathbf{L} = \begin{cases} 116 * Y^{\frac{1}{3}} - 16, & \text{if Y} > 0.008856 \\ 903.3 * Y, & \text{if Y} \leq 0.008856 \end{cases}$ **a = 500(f(X) – f(Y)) + delta** **b = 200(f(Y) – f(Z)) + delta** **where** $\mathbf{f(t)} = \begin{cases} t^{1/3} & \text{for t} > 0.008856 \\ 7.787t + 16/116 & \text{for t} \leq 0.008856 \end{cases}$ **and** $\mathbf{delta} = \begin{cases} 128 & \text{for 8 - bit images} \\ 0 & \text{for floating - point images} \end{cases}$

Table 2.3 Color Feature Calculations

Algorithm	Equation Description
Average	$\boldsymbol{average}\ \mu = \frac{\sum x}{n}$, **n** = blob mask area, **x** = pixel value, e.g. Red in (**R, G, B**)
Standard Deviation	$\boldsymbol{standard\ deviation}\ \sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}}$, **n** = blob mask area, **x** = pixel value, e.g. Red in (**R, G, B**) **μ =** average of feature value
Sine Value	$\boldsymbol{Sine\ value} = x * \frac{opposite}{hypotenuse}$, **x** = hue average value
Cosine Value	$\boldsymbol{Cosine\ value} = x * \frac{adjacent}{hypotenuse}$, **x** = hue average value
Angular Dispersion	$\boldsymbol{Angular\ Dispersion} = \sqrt{\boldsymbol{Sine\ value}^2 + \boldsymbol{Cosine\ value}^2}$

2.4.2. Shape Features. Shape features of the blob masks are also calculated and saved in a csv file. The shape features are shown in Table 2.4. Through observation, SH contains pouch-like white areas; on the other hand, BCC often contains squashed and irregularly shaped white areas. By using the blob masks, the white area regions are segmented out of the lesion image, enabling shape feature extraction. Calculating the shape features of each blob provides additional information to increase classification accuracy. Shape features calculated are: irregularity average, irregularity standard deviation, eccentricity average, and eccentricity standard deviation, shown in Table 2.5. In addition, blotch features previously implemented in research for blotch detection in dermoscopy images [7] were included: eccentricity of the largest blotch, relative size of all blotches, relative size of largest blotch, number of blotches and irregularity of largest blotch. Formulas used for blotch calculations are shown in Table 2.6.

Table 2.4 Shape Features

Feature
Irregularity Average
Eccentricity Average
Number of Blotches
Largest Blotch Eccentricity
Largest Blotch Relative Size
Largest Blotch Irregularity
Relative Size of All Blotches
Eccentricity Standard Deviation
Irregularity Standard Deviation

Table 2.5 Shape Feature Calculations I

Algorithm	Equation Description
Irregularity	$$\boldsymbol{Irregularity} = \frac{\boldsymbol{Perimeter}}{\sqrt{\boldsymbol{Area}}}$$ **Perimeter** = perimeter of blotch **area** = area of blotch
Eccentricity	$$\boldsymbol{Eccentricity} = \frac{\boldsymbol{Minor\ Axis}}{\boldsymbol{Major\ Axis}}$$ **Minor Axis** = minor axis length of the fitted ellipse of blotch **Major Axis** = major axis length of the fitted ellipse of blotch
Average	$\boldsymbol{average}\ \mu = \frac{\sum \boldsymbol{x}}{\boldsymbol{n}}$, **n** = blob mask area, **x** = feature value, e.g. Eccentricity of blotch
Standard Deviation	$\boldsymbol{standard\ deviation}\ \boldsymbol{\sigma} = \sqrt{\frac{\sum(\boldsymbol{x}-\mu)^2}{\boldsymbol{n}}}$, **n** = blob mask area, **x** = feature value, e.g. Eccentricity of blotch **μ** = average of feature value

Table 2.6 Shape Feature Calculations II [7]

Algorithm	Equation Description
Eccentricity of Largest Blotch	$$\textbf{\textit{Eccentricity of Largest Blotch}} = \frac{\boldsymbol{D}}{\sqrt{\boldsymbol{A}}}$$ **D =** distance between the center of largest blotch and the center of lesion **A =** area of lesion
Relative Size of All Blotches	$$\textbf{\textit{Relative Size}} = \frac{\sum \mathbf{Bi}}{\boldsymbol{A}}$$ **Bi** = area of each blotch **A** = area of lesion
Relative Size of Largest Blotch	$$\textbf{\textit{Relative Size of Largest Blotch}} = \frac{\boldsymbol{Bmax}}{\boldsymbol{A}}$$ **Bmax** = area of largest blotch **A** = area of lesion
Number of Blotches	The number of blotches in image
Irregularity of Largest Blotch	$$\textbf{\textit{Irregularity of Largest Blotch}} = \frac{\boldsymbol{Pmax}}{\sqrt{\boldsymbol{Bmax}}}$$ **Pmax** = perimeter of largest blotch **Bmax** = area of largest blotch

3. FEATURE SELECTION

After extracting features from using the blob masks and the features from the lesion area, significant features were sought to distinguish BCC and SH. These significant features were needed to develop the classification training model. In this research, the logistic regression model was used for selecting significant features from among the 51 (21 blob color features, 21 lesion color features, and 9 blob shape features) and determining an intercept and the proper weights for each corresponding feature in the zeta function for the logistic regression model.

3.1. SAS

SAS is a statistic analysis program developed by SAS (Statistic Analytic Software) (SAS Institute, Cary, NC). SAS is used here in the project for selecting significant features from the extracted feature set and for generating a logistic regression model to classify BCC and SH.

3.2. MODEL GENERATION AND FEATURE SELECTION

The logistic regression model works as follows: The stepwise forward option is used to evaluate how well the features perform when introduced one by one to the model. Each feature that remains within the model is evaluated. As a result, the model is constructed by an intercept $\boldsymbol{\beta_0}$, significant features and corresponding weights that were chosen and calculated by the model. The features that were chosen are $\mathbf{x = [x_1\ x_2\ x_3 \ldots\ x_N]}$ and the corresponding weights are $\boldsymbol{\beta = [\beta_1\ \beta_2\ \beta_3 \ldots\ \beta_N]}$. Using these variables and parameters, the formula for calculating the value for classification is

$$f(x) = \frac{1}{(1+ e^{-z(x)})},$$

where

$$\mathbf{z(x) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \ldots + \beta_N * x_N.}$$

After choosing a suitable threshold value T, any value of f(x) above or equal to T is classified as type 1 and any value below T is classified as type 0. Then, the results from the model can then be classified as type 1 which is BCC or type 0 which is SH.

4. RESULTS

After image processing, feature extraction and model generation stages, evaluating the model by implementing the model on the test set is crucial for testing the reliability and accuracy of the model. The following shows the results and evaluations done.

4.1. MODEL GENERATION RESULTS

Eight different data set results are generated by selecting four different area filters and the option of adding or not adding overall lesion color features to the feature sets. The eight data sets are shown in Table 4.1.

Table 4.1 Data Sets Used

Data Set	Area filter (%)	With / Without Overall Lesion Data
Data Set 1	**0.15**	**Without**
Data Set 2	**0.10**	**Without**
Data Set 3	**0.06**	**Without**
Data Set 4	**0.02**	**Without**
Data Set 5	**0.15**	**With**
Data Set 6	**0.10**	**With**
Data Set 7	**0.06**	**With**
Data Set 8	**0.02**	**With**

The models are generated from each training data set using the feature selection procedure in Section 3. The area under the Receiver Operating Characteristic (ROC) curve value and the features that were chosen for each model are presented in Table 4.2.

Table 4.2 Area Under Curve and Features Chosen

Data Set	Area Under Curve (AUC)	Number of Features Chosen	Features Chosen
Data Set 1	**0.916**	**4**	**3, 6, 7, 8**
Data Set 2	**0.905**	**4**	**5, 6, 7, 8**
Data Set 3	**0.914**	**4**	**5, 6, 7, 8**
Data Set 4	**0.903**	**3**	**3, 6, 9**
Data Set 5	**0.953**	**6**	**3, 7, 8, 10, 13, 15**
Data Set 6	**0.959**	**8**	**3, 4, 6, 7, 8, 10, 13, 15**
Data Set 7	**0.965**	**8**	**3, 4, 6, 7, 8, 10, 13, 15**
Data Set 8	**0.968**	**10**	**1, 2, 8, 9, 10, 11, 12, 13, 14, 15**

The features that were chosen and the corresponding names that each feature label represents are shown in Table 4.3.

Table 4.3 Features Chosen

Label	Feature Name
Feature 1	**Red Standard Deviation of Areas Within Blob Mask**
Feature 2	**Hue Standard Deviation of Areas Within Blob Mask**
Feature 3	**Value Standard Deviation of Areas Within Blob Mask**
Feature 4	**L Standard Deviation of Areas Within Blob Mask**
Feature 5	**a* Standard Deviation of Areas Within Blob Mask**
Feature 6	**b* Standard Deviation of Areas Within Blob Mask**
Feature 7	**Irregularity Average of Areas Within Blob Mask**
Feature 8	**Eccentricity Average of Areas Within Blob Mask**
Feature 9	**Irregularity Standard Deviation of Areas Within Blob Mask**
Feature 10	**Green Average of Area Within Lesion Border**
Feature 11	**Hue Average of Area Within Lesion Border**
Feature 12	**Value Average of Area Within Lesion Border**
Feature 13	**Saturation Standard Deviation of Area Within Lesion Border**
Feature 14	**L Average of Area Within Lesion Border**
Feature 15	**L Standard Deviation of Area Within Lesion Border**

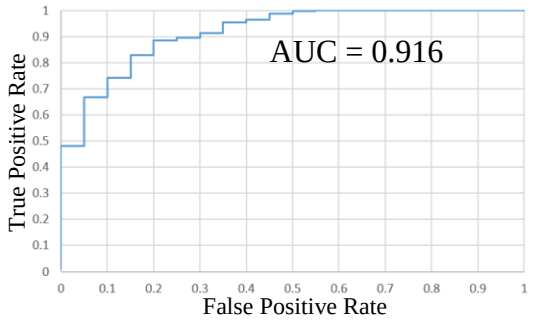
4.2. EVALUATIONS ON MODELS AND TEST RESULTS

After generating the model and running the model on the test set for each data set, accuracy, precision, recall and f-measure mean are calculated for both the training and test sets. The measurement values and the corresponding formulas are shown in Table 4.4 and the thresholds T used based on different sensitivity and specificity of each model are shown in Table 4.5. The number of BCC and SH that were misclassified in the training set and test set of each dataset for each corresponding threshold are shown in Tables 4.6-4.9. The results of the calculations for the eight data set training models are shown in Table 4.10-4.13 for each corresponding threshold, and the results of the calculations for the eight data test sets are shown in Table 4.14-4.17 for each corresponding threshold.

Table 4.4 Measurements and Formulas

Measurements	Formula
Number of Positives and Negatives	Number of Positives p = number of basal cell carcinoma Number of Negatives n = number of sebaceous hyperplasia
True Positive	True Positive TP = Classified as basal cell carcinoma and is truly basal cell carcinoma
False Positive	False Positive FP = Classified as basal cell carcinoma but not truly basal cell carcinoma
True Negative	True Negative TN = Classified as sebaceous hyperplasia and is truly sebaceous hyperplasia
False Negative	False Negative FN = Classified as sebaceous hyperplasia but not truly sebaceous hyperplasia
Accuracy	Accuracy = $\frac{TP+TN}{p+n}$
Precision	Precision = $\frac{TP}{TP+FP}$
Recall	Recall = $\frac{TP}{TP+FN}$
F-measure mean	F-measure mean = $2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision + Recall}}$
Sensitivity	Sensitivity = $\frac{TP}{TP+FN}$
Specificity	Specificity = $\frac{TN}{TN+FP}$
True Positive Rate	True Positive Rate = $\frac{TP}{TP+FN}$
False Positive Rate	False Positive Rate = 1 - Specificity

The ROC curve graphs for Data Set 1-Data Set 8 for both the training set and test set is shown in Figure 4.1-Figure 4.8.



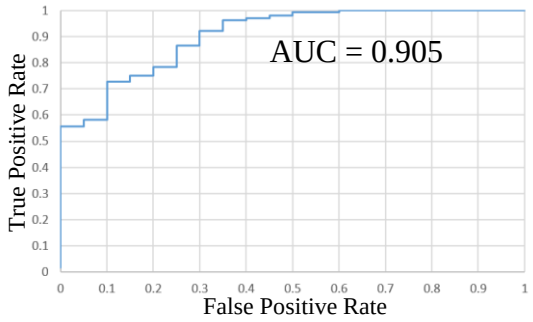
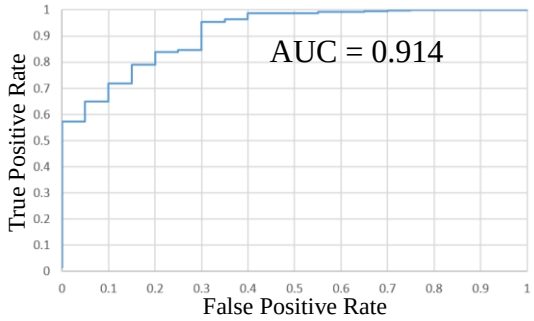


Figure 4.1 Data Set 1 (Left) and Data Set 2 (Right) Training Model ROC Curve



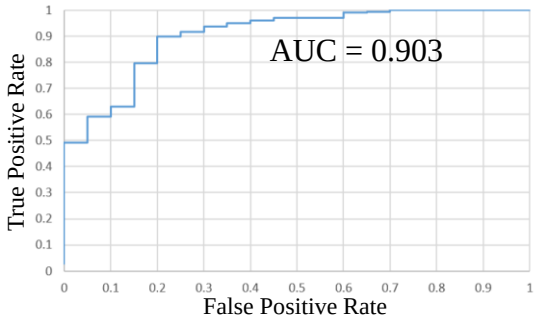
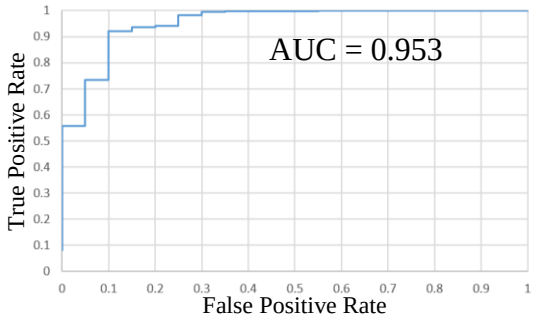


Figure 4.2 Data Set 3 (Left) and Data Set 4 (Right) Training Model ROC Curve



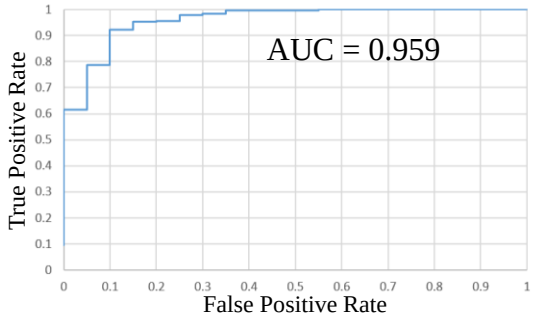
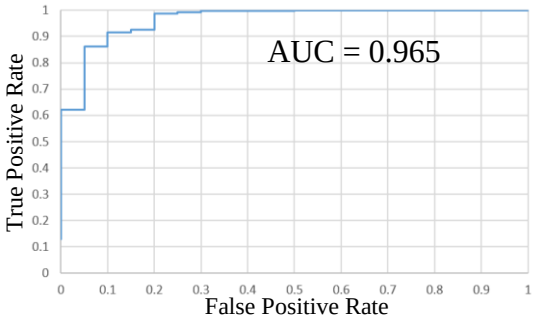


Figure 4.3 Data Set 5 (Left) and Data Set 6 (Right) Training Model ROC Curve



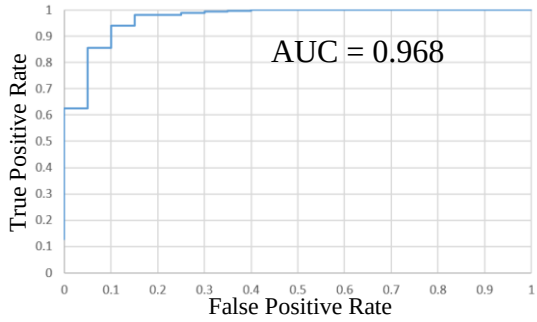
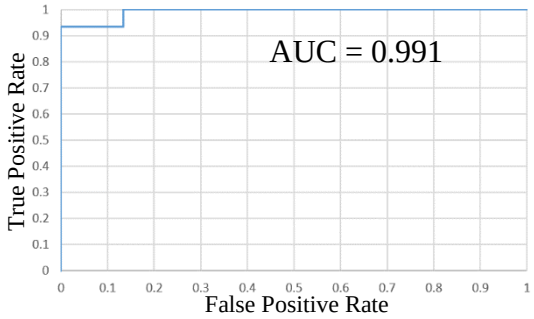


Figure 4.4 Data Set 7 (Left) and Data Set 8 (Right) Training Model ROC Curve



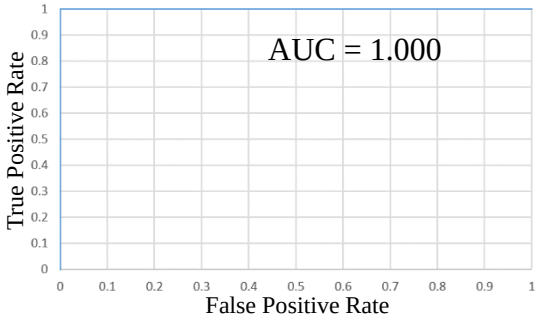
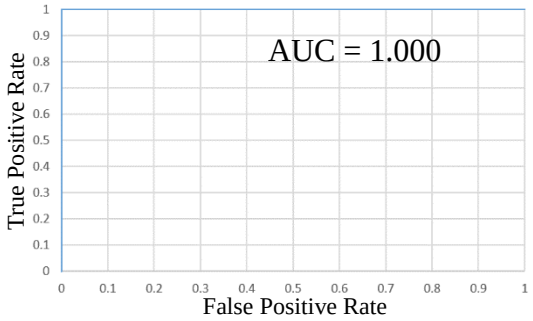


Figure 4.5 Data Set 1 (Left) and Data Set 2 (Right) Test Set ROC Curve



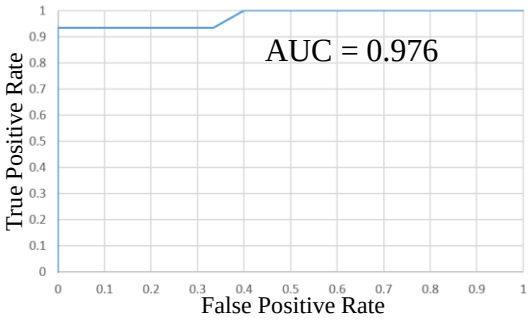
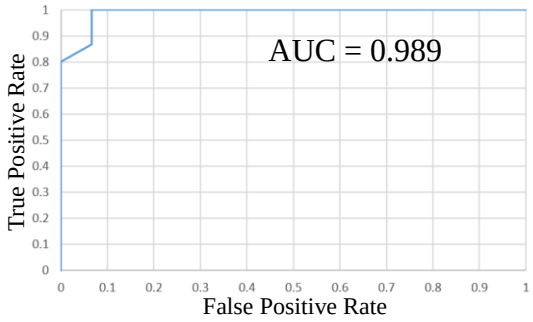


Figure 4.6 Data Set 3 (Left) and Data Set 4 (Right) Test Set ROC Curve



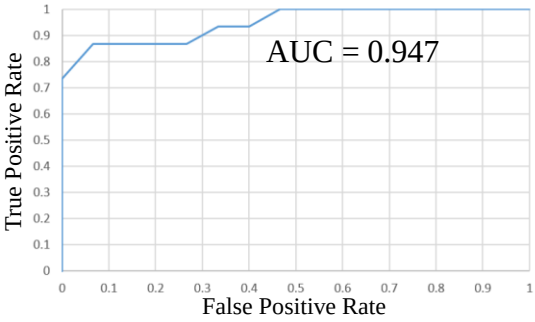
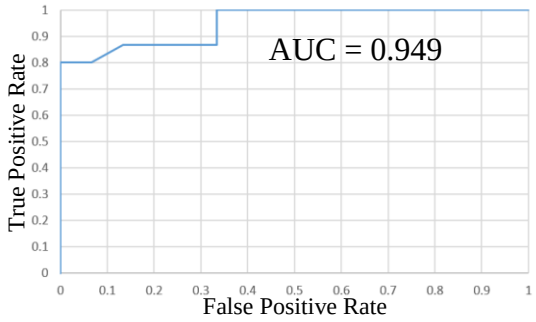


Figure 4.7 Data Set 5 (Left) and Data Set 6 (Right) Test Set ROC Curve



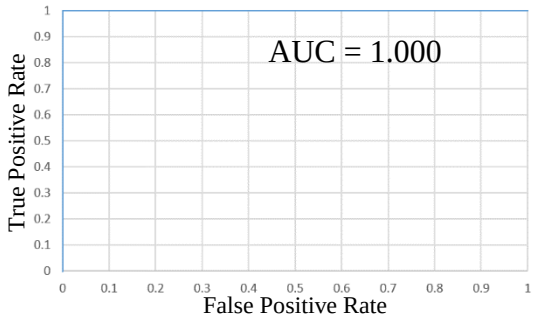


Figure 4.8 Data Set 7 (Left) and Data Set 8 (Right) Test Set ROC Curve

Table 4.5 Thresholds Used for Classification

Threshold	Sensitivity	Specificity
Threshold 1	**0.99**	**high**
Threshold 2	**0.98**	**high**
Threshold 3	**high**	**0.60**
Threshold 4	**high**	**0.70**

The value “high” represents the highest value chosen while maintaining the other column’s chosen value. For example, for Threshold 1, Specificity is chosen as high as possible in the model, allowing Sensitivity to maintain a value closest to 0.99.

Table 4.6 Number of Misclassified Images for Threshold 1

Data Set	Training Set		Test Set	
	Basal Cell Carcinoma	Sebaceous Hyperplasia	Basal Cell Carcinoma	Sebaceous Hyperplasia
Data Set 1	**2**	**10**	**0**	**2**
Data Set 2	**2**	**10**	**0**	**0**
Data Set 3	**3**	**12**	**0**	**3**
Data Set 4	**3**	**12**	**0**	**7**
Data Set 5	**4**	**6**	**2**	**1**
Data Set 6	**1**	**7**	**2**	**2**
Data Set 7	**3**	**6**	**2**	**3**
Data Set 8	**1**	**7**	**0**	**2**

Table 4.7 Number of Misclassified Images for Threshold 2

Data Set	Training Set		Test Set	
	Basal Cell Carcinoma	Sebaceous Hyperplasia	Basal Cell Carcinoma	Sebaceous Hyperplasia
Data Set 1	6	9	1	1
Data Set 2	6	7	0	0
Data Set 3	4	9	0	0
Data Set 4	4	12	1	5
Data Set 5	6	6	3	0
Data Set 6	5	6	2	1
Data Set 7	5	4	2	2
Data Set 8	6	5	0	1

Table 4.8 Number of Misclassified Images for Threshold 3

Data Set	Training Set		Test Set	
	Basal Cell Carcinoma	Sebaceous Hyperplasia	Basal Cell Carcinoma	Sebaceous Hyperplasia
Data Set 1	14	8	2	0
Data Set 2	12	8	0	0
Data Set 3	11	8	0	0
Data Set 4	14	8	1	2
Data Set 5	1	7	0	3
Data Set 6	1	8	1	5
Data Set 7	1	8	2	4
Data Set 8	1	8	0	2

Table 4.9 Number of Misclassified Images for Threshold 4

Data Set	Training Set		Test Set	
	Basal Cell Carcinoma	Sebaceous Hyperplasia	Basal Cell Carcinoma	Sebaceous Hyperplasia
Data Set 1	35	5	3	0
Data Set 2	34	6	3	0
Data Set 3	19	6	0	0
Data Set 4	29	5	1	1
Data Set 5	2	6	0	1
Data Set 6	17	3	4	0
Data Set 7	4	4	2	2
Data Set 8	8	3	0	0

Table 4.10 Results of Measurements of Training Model for Threshold 1

Data Set	Accuracy	Precision	Recall	F-measure mean
Data Set 1	0.9660	0.970674	0.993994	0.982196
Data Set 2	0.9660	0.970674	0.993994	0.982196
Data Set 3	0.9575	0.964912	0.990991	0.977778
Data Set 4	0.9575	0.964912	0.990991	0.977778
Data Set 5	0.9717	0.982090	0.987988	0.985030
Data Set 6	0.9773	0.979351	0.996997	0.988095
Data Set 7	0.9745	0.982143	0.990991	0.986547
Data Set 8	0.9773	0.979351	0.996997	0.988095

Table 4.11 Results of Measurements of Training Model for Threshold 2

Data Set	Accuracy	Precision	Recall	F-measure mean
Data Set 1	0.9575	0.973214	0.981982	0.977578
Data Set 2	0.9547	0.979042	0.981982	0.980510
Data Set 3	0.9632	0.973373	0.987988	0.980626
Data Set 4	0.9547	0.964809	0.987988	0.976261
Data Set 5	0.9660	0.981982	0.981982	0.981982
Data Set 6	0.9688	0.982036	0.984985	0.983508
Data Set 7	0.9745	0.987952	0.984985	0.986466
Data Set 8	0.9688	0.984940	0.981982	0.983459

Table 4.12 Results of Measurements of Training Model for Threshold 3

Data Set	Accuracy	Precision	Recall	F-measure mean
Data Set 1	0.9377	0.975535	0.957958	0.966667
Data Set 2	0.9433	0.975684	0.963964	0.969789
Data Set 3	0.9462	0.975758	0.966967	0.971342
Data Set 4	0.9377	0.975535	0.957958	0.966667
Data Set 5	0.9773	0.979351	0.996997	0.988095
Data Set 6	0.9745	0.976471	0.996997	0.986627
Data Set 7	0.9745	0.976471	0.996997	0.986627
Data Set 8	0.9745	0.976471	0.996997	0.986627

Table 4.13 Results of Measurements of Training Model for Threshold 4

Data Set	Accuracy	Precision	Recall	F-measure mean
Data Set 1	0.8867	0.983498	0.894895	0.937107
Data Set 2	0.8867	0.980328	0.897898	0.937304
Data Set 3	0.9292	0.981250	0.942943	0.961715
Data Set 4	0.9037	0.983819	0.912913	0.947040
Data Set 5	0.9773	0.982196	0.993994	0.988060
Data Set 6	0.9433	0.990596	0.948949	0.969325
Data Set 7	0.9773	0.987988	0.987988	0.987988
Data Set 8	0.9688	0.990854	0.975976	0.983359

Table 4.14 Results of Measurements of Test Set Results for Threshold 1

Data Set	Accuracy	Precision	Recall	F-measure mean
Data Set 1	0.9333	0.882353	1.000000	0.937500
Data Set 2	1.0000	1.000000	1.000000	1.000000
Data Set 3	0.9000	0.833333	1.000000	0.909091
Data Set 4	0.7667	0.681818	1.000000	0.810811
Data Set 5	0.9000	0.928571	0.866667	0.896552
Data Set 6	0.8667	0.866667	0.866667	0.866667
Data Set 7	0.8333	0.812500	0.866667	0.838710
Data Set 8	0.9333	0.882353	1.000000	0.937500

Table 4.15 Results of Measurements of Test Set Results for Threshold 2

Data Set	Accuracy	Precision	Recall	F-measure mean
Data Set 1	0.9333	0.933333	0.933333	0.933333
Data Set 2	1.0000	1.000000	1.000000	1.000000
Data Set 3	1.0000	1.000000	1.000000	1.000000
Data Set 4	0.8000	0.736842	0.933333	0.823529
Data Set 5	0.9000	1.000000	0.800000	0.888889
Data Set 6	0.9000	0.928571	0.866667	0.896552
Data Set 7	0.8667	0.866667	0.866667	0.866667
Data Set 8	0.9667	0.937500	1.000000	0.967742

Table 4.16 Results of Measurements of Test Set Results for Threshold 3

Data Set	Accuracy	Precision	Recall	F-measure mean
Data Set 1	0.9333	1.000000	0.866667	0.928571
Data Set 2	1.0000	1.000000	1.000000	1.000000
Data Set 3	1.0000	1.000000	1.000000	1.000000
Data Set 4	0.9000	0.875000	0.933333	0.903226
Data Set 5	0.9000	0.833333	1.000000	0.909091
Data Set 6	0.8000	0.736842	0.933333	0.823529
Data Set 7	0.8000	0.764706	0.866667	0.812500
Data Set 8	0.9333	0.882353	1.000000	0.937500

Table 4.17 Results of Measurements of Test Set Results for Threshold 4

Data Set	Accuracy	Precision	Recall	F-measure mean
Data Set 1	**0.9000**	**1.000000**	**0.800000**	**0.888889**
Data Set 2	**0.9000**	**1.000000**	**0.800000**	**0.888889**
Data Set 3	**1.0000**	**1.000000**	**1.000000**	**1.000000**
Data Set 4	**0.9333**	**0.933333**	**0.933333**	**0.933333**
Data Set 5	**0.9667**	**0.937500**	**1.000000**	**0.967742**
Data Set 6	**0.8667**	**1.000000**	**0.733333**	**0.846154**
Data Set 7	**0.8667**	**0.866667**	**0.866667**	**0.866667**
Data Set 8	**1.0000**	**1.000000**	**1.000000**	**1.000000**

5. CONCLUSION

5.1. RESULTS AND DISCUSSIONS

From comparing the thresholds of all four measurement results tables for the training set, in Tables 4.10-4.13, Threshold 1 has the overall highest F-measure mean and has the highest individual F-measure mean among all the results for Threshold 1-Threshold 4. Using Threshold 1, Data Set 6 and Data Set 8 had the highest F-measure means. If we compare the two data sets, the two sets differ in the features chosen for the models and the area under the curve. Even though Data Set 8 had a higher value of area under the curve, the number of features chosen for the model is higher than for Data Set 6. Therefore, regarding successful classification, the two sets are similar.

In comparing all eight data sets, Data Set 5 using Threshold 3, Data Set 6 using Threshold 1 and Data Set 8 using Threshold 1 all had the highest F-measure mean value, and all three sets missed one basal cell carcinoma image and seven sebaceous hyperplasia images. We may further compare the three sets regarding all four thresholds. Even though Threshold 1 in Data Set 5 had a worse F-measure mean than Data Set 6 and Data Set 8, all three sets have good results for Thresholds 2 and 3. For Threshold 4, the F-measure mean for Data Set 5 is better than for Data Sets 6 and 8.

5.2. LIMITATIONS AND FUTURE WORK

The limitation of this project is having an unbalanced training set, where there are 333 BCC images and 20 SH images. Hence, if we were to estimate the accuracy of BCC detection for all images in the training set, the accuracy would be 94.33% if we were to guess BCC for all images. Even though the training set is unbalanced, only 8 out of 32 results for training accuracy are $\leq$ 94.33%. Moreover, all training set accuracy results for Threshold 1 and Threshold 2 are > 94.33%. On the other hand, the test set is not biased, where there are 15 BCC images and 15 SH images. Hence, if we were to guess BCC for all images in the test set, the accuracy would be 50.00%. Whereas, the test set accuracy from all 32 results > 76.00%, exceeding by 26.00% the accuracy obtained guessing all BCC.

Even though the training set is unbalanced, from the results and discussion above, the methodology used in the project can increase the discrimination of BCC dermoscopy

images from SH dermoscopy images. For future work, the next step will be to obtain more SH dermoscopy images until the number reaches a number close to the number of BCC dermoscopy images to obtain a balanced training and test set. Another step would be to balance the lighting within the lesion images because with some areas lighter or darker than other areas, several white areas were merged together and some were removed during the thresholding image processing stage. With balanced image lighting, segmentation of white areas is expected to improve. Furthermore, recent research on image recognition has shown that the implementation of deep learning methods, especially convolution neural networks, has the tendency to obtain great results for classification. Therefore, the fusion of deep learning methods with the project is expected to improve the classification abilities of the models.

BIBLIOGRAPHY

[1] S. Coulter, "Basal Cell Carcinoma," *http://sarahcoulter.myefolio.com/Uploads/Basal%20Cell%20Carcinoma.pdf*, 2014.

[2] H. W. Rogers, M. A. Weinstock, S. R. Feldman. B. M. Coldiron, "Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the US population, 2012," *JAMA Dermatol*, vol. 151, no.10, pp. 1081–1086, 2015.

[3] E. Bisaccia, K. Lee and D. Scarborough, "Treatment of Extensive Sebaceous Hyperplasia," *the dermatologist*, vol. 19, issue 4, 2011.

[4] A. Ayoub, A. Hajdu and A. Nagy, "Automatic Detection of Pigmented Network in Melanoma Dermoscopic Images," *The International Journal of Computer Science and Communication Security(IJCSCS)*, vol. 2, pp. 58-63, 2012.

[5] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-9, no. 1, pp. 62-66, 1979.

[6] OpenCV Documentations, *http://docs.opencv.org/2.4/modules/imgproc/doc/miscellaneous_transformations.html#cvtcolor*

[7] W. V. Stoecker, K. Gupta, R. J. Stanley, R. H. Moss, and B. Shrestha, "Detection of Asymmetric Blotches (Asymmetric Structureless Areas) in Dermoscopy Images of Malignant Melanoma Using Relative Color," *Skin Research and Technology*, vol. 11, no. 3, pp. 179-184, 2005.

[8] R. Kaur, R. LeAnder, N. K. Mishra, J. R. Hagerty, R. Kasmi, R. J. Stanley, M. E. Celebi and W. V. Stoecker, "Thresholding Methods for Lesion Segmentation of Basal Cell Carcinoma in Dermoscopy Images," *Skin Research and Technology*, in press.

[9] A. Banerjee, "Impact of Principal Component Analysis in the Application of Image Processing," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, issue 1, 2012.

VITA

Jonathan Liao was born in the United States. He earned his Bachelor of Science degree in Electrical Engineering from National Taipei University of Technology, Taipei, Taiwan in May 2015. He earned his Master of Science degree in Electrical Engineering from Missouri University of Science and Technology, Rolla, Missouri, USA in May 2017.