Spring 2017

Classification of basal cell carcinoma using telangiectatic vessels and machine learning

Hemanth Yadav Aradhyula

Department:

CLASSIFICATION OF BASAL CELL CARCINOMA USING TELANGIECTATIC VESSELS AND MACHINE LEARNING

by

HEMANTH YADAV ARADHYULA

A THESIS

Presented to the Graduate Faculty 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. Bijaya Shrestha
Dr. R. Joe Stanley

ABSTRACT

Basal cell carcinoma (BCC) is one of the most common types of skin cancer in the United States. Early detection of BCC by noninvasive techniques can decrease delay in treatment and save cost. A recent study estimated that 5.4 million cases of non-melanocytic skin cancer (NMSC) occur each year in the US. BCC accounts for 50% of NMSC cases. Telangiectasia, which appears in most BCCs is an important feature for identification of BCC for an automatic diagnostic system. In this thesis, three methods for detection of telangiectasia present in dermoscopy lesion image (DI) were proposed. Detected telangiectasia in DI was used to predict BCC. Using stepwise logistic regression, a model was created for which the area under a receiver operating characteristic (ROC) curve of 88.9% was achieved for detection of BCC.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Randy Moss, for his support throughout my master's program. He is one of the best persons I have ever meet and a great mentor. His ability to explain and patience are paramount. I would like to thank Dr. Stoecker for guiding me through a very unfamiliar topic. His knowledge is impeccable. I am very grateful for Dr. Randy Moss and Dr. Stoecker for sharing their immense knowledge graciously all through my research. I would thank my other committee members, Dr. R Joe Stanley and Dr. Bijaya Shrestha, for their support and feedback. All of your commitment and passion for research is an inspiration and stays with me forevery.

I would like to thank Nabin Mishra and Jason Hagerty for their ideas and assistance. They are some of the most talented people I have worked with. I appreciate the efforts of Sadhika Jagannathan, Daniyal Saeed, and Reda Kasmi. I would thank my parents, family, and friends for helping me make right decisions and for standing by my side whenever I need them. I would also thank all the people who supported me in any way for my research.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

LIST OF ABBREVIATIONS

ACC accuracy. xii, 38, 43

ALB automatic lesion borders. xii, 2, 3

AUC area under ROC curve. xii, 36, 37

BAN brown area network. xii, 8, 10, 12, 20–25

BCC basal cell carcinoma. iii, x, xii, 1–3, 6, 26–28, 32, 34–37, 43–46

CD color drop filter. xii, 8, 20, 22, 24

CM Cheng Model. xii, 1, 36, 46

DI dermoscopy lesion image. iii, viii, xii, 1–5, 10, 14, 21, 23, 25, 31

FEA features. xii, 38, 43

fscore F-score. xii

HMI hair mask images. xii, 2

INT feature interaction up to two. xii, 38, 43

MM1 Modified Model 1. xii, 22, 24, 36, 46

MM2 Modified Model 2. viii, xii, 24, 25, 36, 46

MVM manual vessel mask. xii, 2

NMSC non-melanocytic skin cancer. iii, xii, 1

NonBCC benign dermoscopy image. xii, 2, 3, 8, 34

PPV Positive predictive value. xii

RCF red chromaticity filter. xii, 10, 11, 22–25

ROC receiver operating characteristic. iii, xii, 36, 37

SC squamous cell carcinomas. xii, 1

SCD second color drop filter. xii, 24, 25

SEN sensitivity. xii, 38, 43

SLE SLENTRY. xii, 36–38, 43

SLS SLSTAY. xii, 36–38, 43

SPC specificity. xii, 38, 43

VM vessel mask. viii, xii, 4, 7–27, 29–31, 35, 46, 47

WAF white area filter. viii, xii, 11, 22, 24

GLOSSARY

chromaticity an objective specification of the quality of a color regardless of its luminance. viii, xii, 9, 10

eccentricity the ratio of the distance between the foci of the ellipse and its major axis length. xii, 27, 30, 31

obs observed Bcc and non NonBCC Images . xii

pre predicated Bcc and non NonBCC Images. xii

skin blue an average of the blue values outside of Lesion. xii, 12, 26

skin green an average of the green values outside of Lesion. xii, 12, 26

skin red an average of the red values outside of Lesion. xii, 8, 12, 13, 26

telangiectasia small dilated blood vessels near the surface of the skin. iii, xii, 1, 6, 7, 20, 46, 47

1. INTRODUCTION

BCC is one of the most common nonmalignant skin cancers in the USA. According to American Cancer Society, 5.4 million cases of non-melanocytic skin cancer (NMSC) were recorded in 2012 [1] [2]. BCC accounts for 50% of NMSC, while squamous cell carcinomas (SC) make up other 50% [3]. BCC has a very low metastatic risk, but they can cause permanent disfigurement, especially around the eye, nose, and ear [4]. Most BCCs have telangiectasia in their lesion area. Early detection of BCC is crucial in reducing the cost and delay in treatment. Non-melanocytic skin cancer is the fifth most costly cancer due to the sheer volume of cases [5].

Detection of telangiectasia can be very useful in the classification of BCC [6]. Telangiectases are small dilated blood vessels that appear near the skin surface. Clinically, BCC was detected using biopsy of a tissue sample. Noninvasive dermoscopy lesion image (DI) were used for analyses. The focus of this research is to improve the classification accuracy of the work done by Cheng [6]. Two modified models for better detection of telangiectasia were proposed apart from the Cheng Model (CM). In addition, a larger image set with more features was used. A comparison with the Cheng model [6] is also provided.

2. DATA SET

A total of 1024 BCC and NonBCC dermal images of 1024 x 768 pixels were used in the analysis, compared to the 175 dermoscopy lesion image (DI) used by Cheng [6]. DI that were generated using the DermLite HR II with a gel interface on the lesion were used. Apart from BCC and NonBCC, automatic lesion borders (ALB), hair mask images (HMI), and manual vessel masks (MVM), were used for the analysis.

Dermoscopy lesion image used for this analysis were acquired from four clinics located in Plantation, FL; Rolla, MO; Columbia, MO; and Stamford, CT. Only two of the BCCs were detected using confocal microscopy, while others were biopsied and examined by a dermatopathologist.

2.1. BASAL CELL CARCINOMA (BCC) IMAGES

Dermoscopy images were diagnosed with BCC by using biopsy. There are 304 BCC images in the analyzed data set. An example of a BCC dermoscopy lesion image is shown in Figure 2.1.

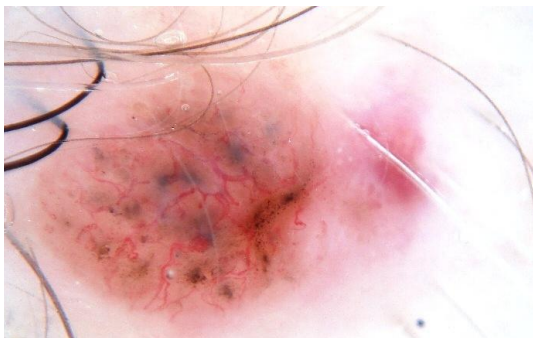


Figure 2.1. Sample BCC dermoscopy images

2.2. BENIGN DERMOSCOPY IMAGES (NONBCC)

Dermoscopy lesion images which were not diagnosed as BCC. Benign images were either detected by biopsy or by using dermoscopy imagesand following the patient to make sure the lesion was unchanged. There are 720 NonBCC images, of which 292 were nevi, 124 were seborrheic keratosis, 89 were dysplastic nevi, 5 were sebaceous hyperplasia, and the others had various benign diagnoses. An example of a NonBCC dermoscopy lesion image is shown in Figure 2.2.

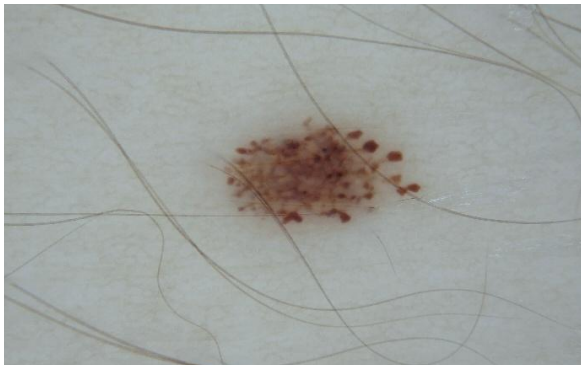
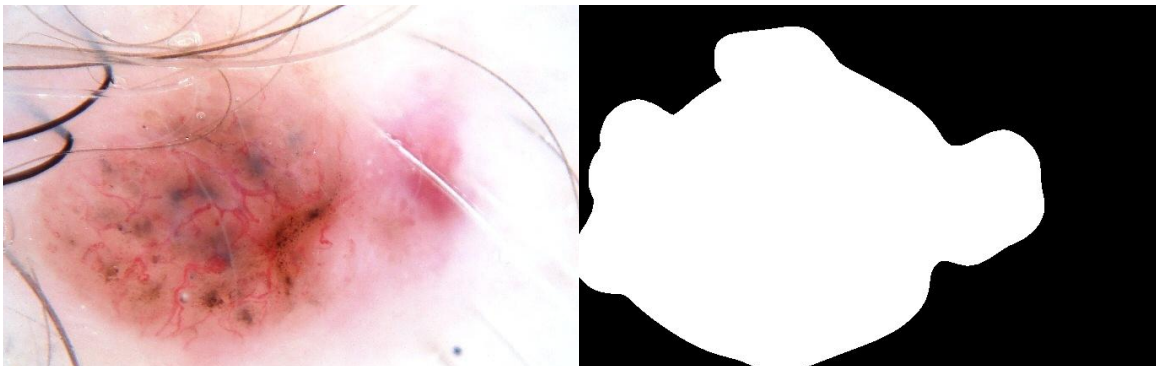


Figure 2.2. Sample Benign dermoscopy image

2.3. AUTOMATIC LESION BORDERS (ALB) IMAGES

Lesion border masks for corresponding BCC and NonBCC images are automatically selected by a classifier out of thirteen different borders generated by using different border algorithms [7]. These borders reduce computational load as analysis algorithms focus on the interior of the lesion border. An example of ALB is provided by Figure 2.3.

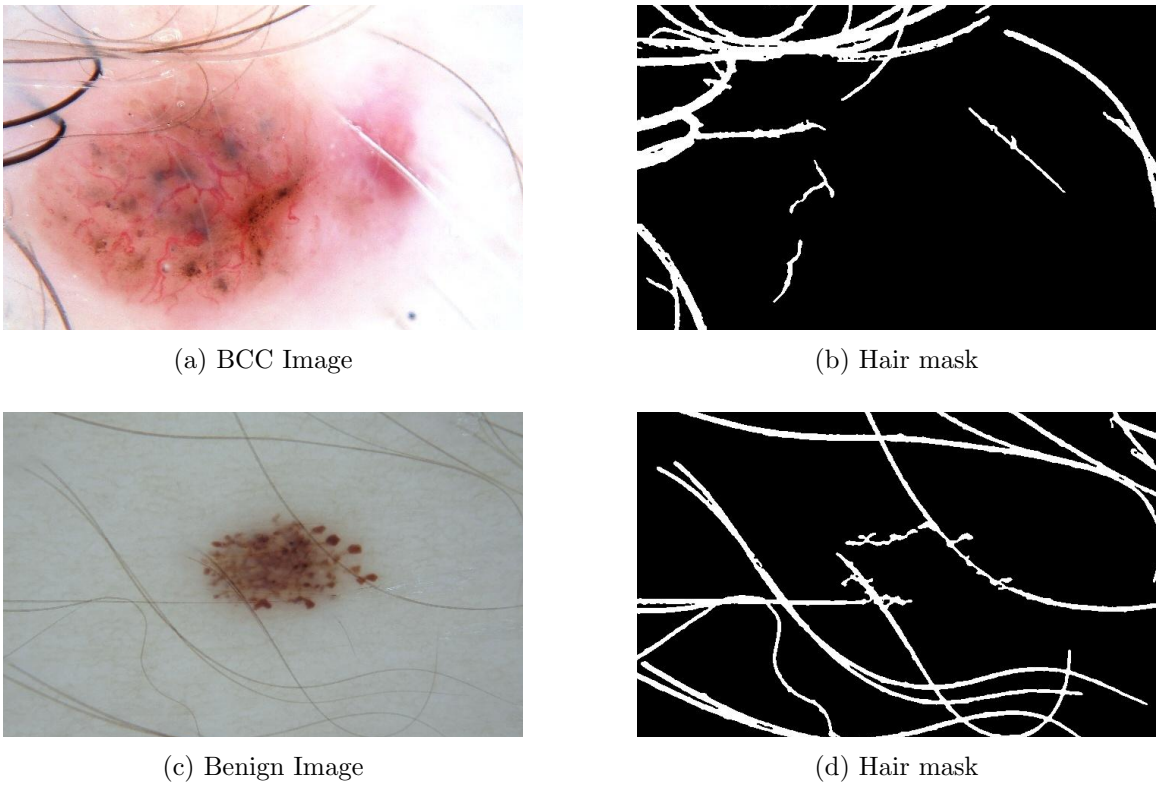


(a) BCC Image (b) Automatic lesion border mask

Figure 2.3. Lesion image for a BCC image

2.4. HAIR MASK IMAGES (HMI)

The lesion area is usually shaved before a dermoscopy image is taken, but it is not uncommon to find hair in the lesion area. Hair masks generated by Reda Kasmi were used to filter the hair from the images [7]. Hair masks for corresponding dermoscopy lesion image are shown in Figure 2.4. Thought the hair mask detects most of the hair, it also detects vessels as hair, as is visible in Figure 2.4b, which results in the loss of some important vessels in the final vessel mask.



(a) BCC Image

(b) Hair mask

(c) Benign Image

(d) Hair mask

Figure 2.4. Hair masks for dermoscopy lesion image

3. TELANGIECTASIA ANALYSIS FILTERS

A telangiectasia is a small dilated blood vessel often detected in the lesion of BCC. Telangiectaes that are narrow and longer are of interest, whereas the wider and shorter most likely occur in sunburn. Different filters were developed to detect telangiectasia in BCC, which are discussed below.

3.1. COLOR-DROP VESSEL DETECTION

Color-drop vessel detection is the main algorithm that is used to detect telangiectasia in the BCC. By visual analysis of Figure 2.1, telangiectasia seems to be red compared to the background, but the color drop from surrounding tissue to the telangiectasia in red(R), green (G), and blue(B) planes reveals that there is a considerable drop in the G and B planes, while the R intensity may increase. The center pixel c shown in Figure 3.1 is slid on every pixel of the image in the lesion. The algorithm iteratively moves one pixel away from the center pixel c in all possible 135 degree angles (for instance N, SE has 135 degrees) and requires a minimum color drop of -2, 7, 4 in R, G, and B planes respectively, until reaching 4 pixels from the center pixel c. For (N,SE), if any pixel combination from the center c (N1,SE1 or N2,SE2 or N3,SE3 or N4,SE4) satisfies the minimum drop then c is marked as a vessel pixel. The color-drop vessel detector is very good at detecting vessels in most of the images, but with present drop values for some images, vessels may be detected as broken vessels instead of a single vessel. The green color in Figure 3.2b represents a vessel mask after the colordrop filter is applied.

Clearly, from Figure 3.2b, it can be seen that all the telangiectasia vessels were found along with a lot of noise. The next filter attempts to remove this extra noise. Pixels that satisfy the above mentioned color drop criteria are used to create a vessel mask.

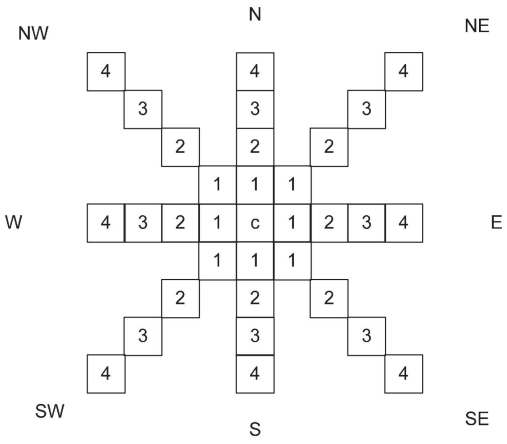
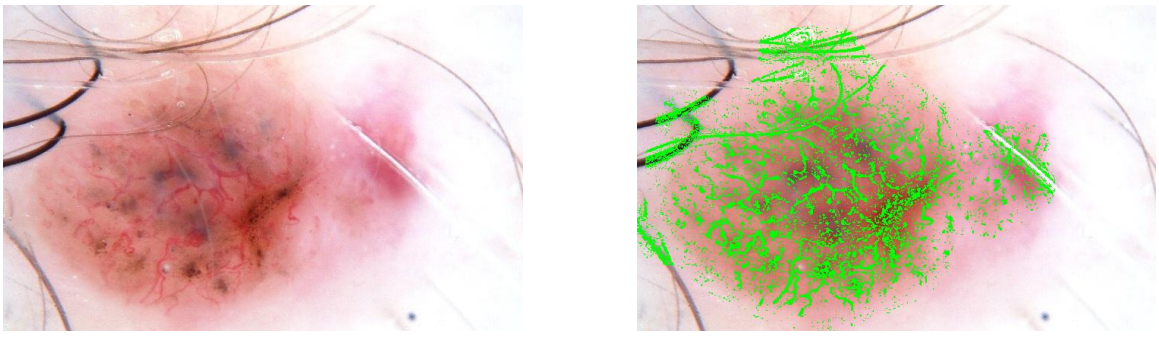


Figure 3.1. Direction mask

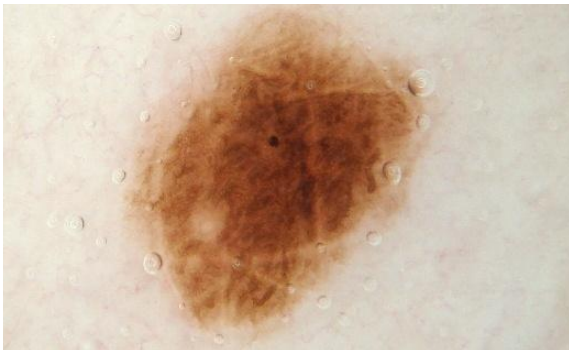


(a) BCC image (b) Color drop image

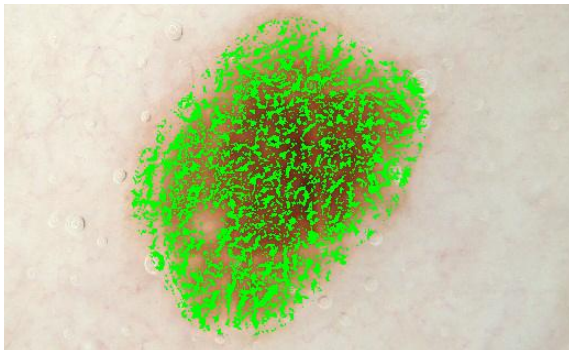
Figure 3.2. Example for performance of Color-drop vessel detection

3.2. BROWN NETWORK FILTER

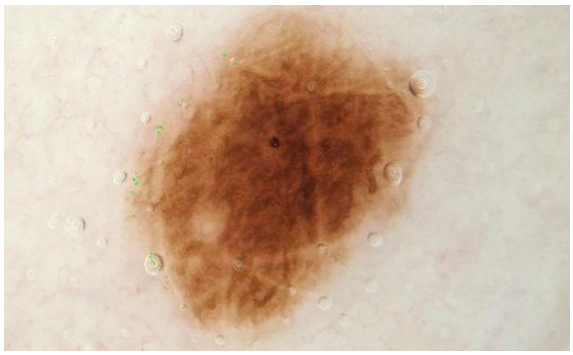
A brown area network is present in the lesion area of many images, especially NonBCC. Figure 3.3a demonstrates one such image, and its corresponding color drop filter image is shown in Figure 3.3b. Let R, G, B represent the pixel intensity value in the three channels and let skin red as an average of the surrounding lesion red value. As brown is degraded orange, every pixel in Figure 3.3b that satisfies the criteria $G > B + 5$ and $R <$ skin red were considered as part of the brown network and thus removed from the vessel mask, which results in Figure 3.3c.



(a) Dermoscopy image



(b) color drop image



(c) Brown network image

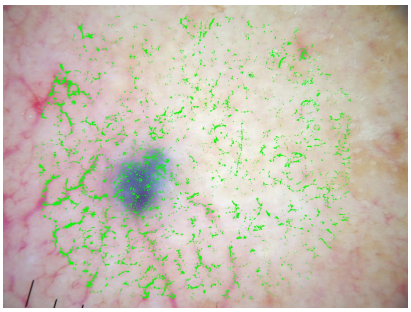
Figure 3.3. Example for brown network filter performance

3.3. RED CHROMATICITY FILTER

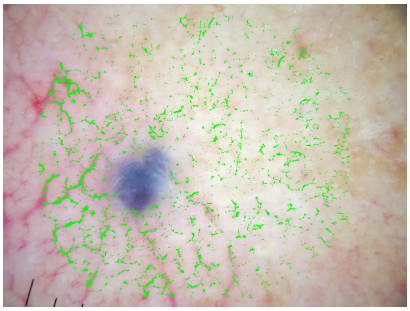
In this filter, the red chromaticity of every pixel inside the lesion area is calculated using equation (3.1).

$$redchromacity = \frac{R}{(R+G+B)} \tag{3.1}$$

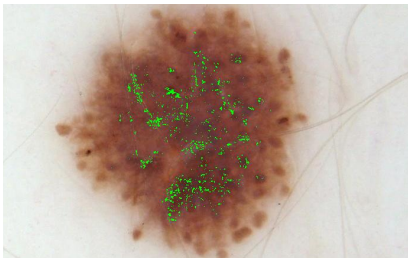
If red chromaticity is greater than 0.33 and less than 0.62 (Eq. 3.2) and the current red intensity is greater than 100 (Eq. 3.3) then the pixel is considered as part of the vessel mask. It is observed that most vessels have 100 as the minimum intensity for red values. This filter reduces dark patches, some brown networks, some dark hair, and some random noise, as shown in the Figure 3.4



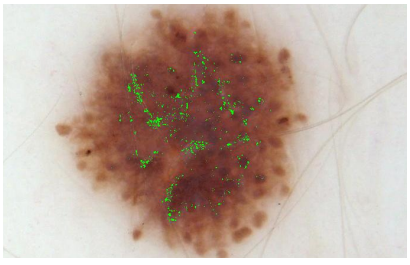
(a) Vessel mask after brown area filter



(b) Red chromaticity filtered image



(c) Vessel mask after brown area filter



(d) Red chromaticity filtered image

Figure 3.4. Vessel mask before and after red chromaticity filter

$$0.33 < \text{redchromacity} < 0.62 \tag{3.2}$$

$$R > 100 \tag{3.3}$$

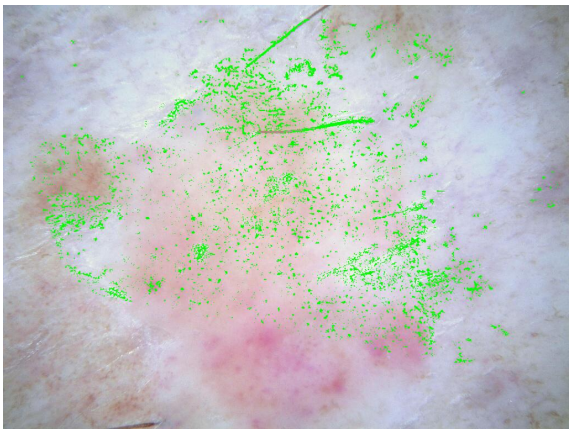
Filtering of noise by the red chromaticity filter is visible from Figure 3.4. Vessel masks obtained from a colordrop filter were passed through the brown area network filter to obtain Figure 3.4a and Figure 3.4c. Figure 3.4b and Figure 3.4d were obtained after filtering the vessels obtained from the brown area network with a red chromaticity filter.

3.4. WHITE AREA FILTER

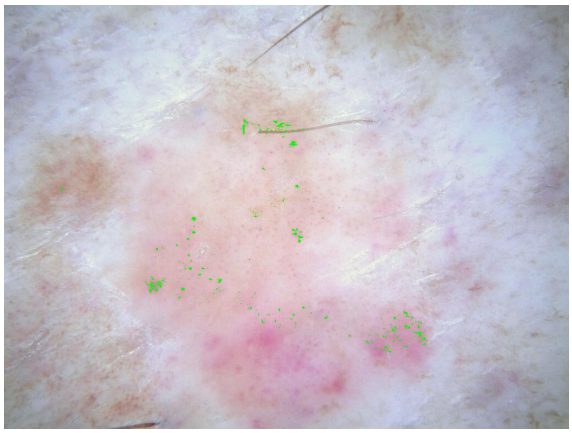
White area noise is formed in the dermoscopy lesion image due to reflection of the light in the gel in addition to noise created due to bubbles in the gel. Let R, G, and B represent red, green, and blue color planes of a color image. Pixels that have R, G, B values greater than 200 intensity or $R < G + 40$ and $R < B + 40$ were removed from the vessel mask (Eq. 3.4):

$$((R > 200) \wedge (G > 200) \wedge (blue > 200)) \vee ((R < G + 40) \wedge (R < B + 40)) \tag{3.4}$$

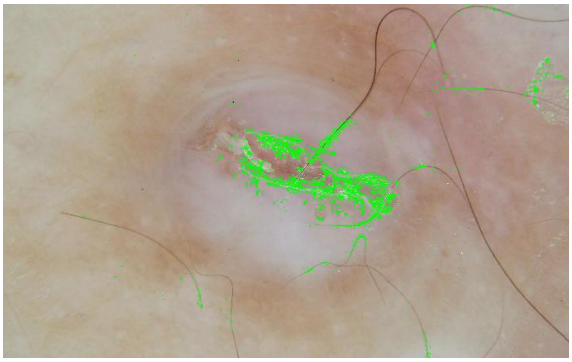
This filter removes white area noise and bubble noise, as seen in Figure 3.5. A vessel mask image obtained after red chromaticity filter is shown in Figure 3.5a and the resultant vessel mask image after white area mask is shown in Figure 3.5b. It can be observed that most of the noise in the white area is removed. Figure 3.5c has noise due to bubbles and white area filter has removed most of the noise as shown in Figure 3.5d.



(a) Vessel mask after red chromaticity filter



(b) Vessel mask after white area filter



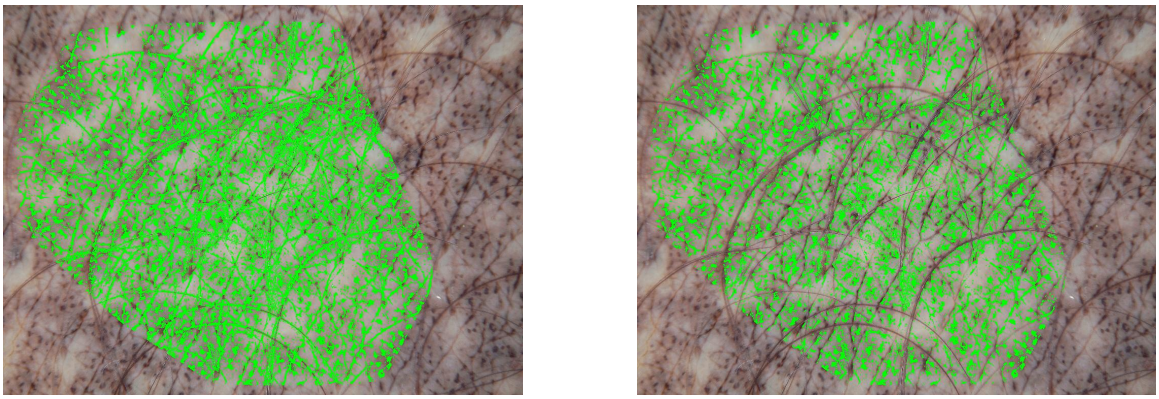
(c) Vessel mask after red chromaticity filter

(d) Vessel mask after white area filter

Figure 3.5. Vessel mask before and after white area filter

3.5. HAIR FILTER

Hair is one of the most prominent noises in the vessel mask. A hair mask was generated by Reda Kasmi [7]. The present version of the hair mask also has some of the vessels detected as hair which results in loss of some of the important vessels in the vessel mask. An example of hair filter is shown in Figure 3.6.



(a) Brown area network filtered image (b) Hair mask filtered image

Figure 3.6. Vessel mask before and after hair mask filter

3.6. BUBBLE FILTER

Bubbles have sharp edges resulting in a steep drop of local intensity compared to the mean intensity of a large area. This effect is mostly visible in the blue plane. The removal of bubbles and other noise is done in three stages. Let R, G, B be the red, green, and blue values of a pixel in the lesion of a dermoscopy lesion image. Skin red, skin green, skin blue are the average red, green, and blue values of the outside lesion area, which is normal skin. Analysis for any selected pixel in the vessel mask is performed on the corresponding dermoscopy pixel.

3.6.1. Stage 1. If $R > 2*G$ and $R > 2*B$ (while * represent multiplication) nothing is changed; if the above condition is false and $B < skinblue - 90$, then the pixel is removed from the vessel mask:

$$\begin{gathered} if(not(R > 2*G \wedge R > 2*B) \wedge (B < \text{skin blue} - 90)) \\ VM(x,y) = 0 \end{gathered} \tag{3.5}$$

3.6.2. Stage 2. If any pixel in the vessel mask has $(B > G) \wedge (G > R*0.7) \wedge$ (skin red $- R > 30) \wedge (R > 140)$ then the pixel in the vessel remains as the pixel

in a vessel. If the above condition is not met, then $R < 130$ should be checked. If $R < 130 \wedge G > R * 0.6$ then the pixel from the vessel should be removed; if $R < 130$ is false, then the pixel should be checked for $R >$ skin red $+ 30$. If $R >$ skin red $+ 30$ is true and $G > R * 0.6$, then the pixel should be removed from the vessel mask. If $R >$ skin red $+ 30$ is false, then the pixel should be checked for $R <=$ skin red. If $R <=$ skin red is true and $G > R * 0.7$, then the pixel should be removed from vessel mask. If $(R <=$ skin red $+ 30) \wedge (R >$ skin red$) \wedge (|G - B| > 10)$ is true, then the pixel should be removed from the vessel mask.

The below equations are used in the flow diagram in Figure 3.7:

$$(B > G) \wedge (G > R * 0.7) \wedge (\text{skin red} - R > 30) \wedge (R > 140) \tag{D1}$$

$$R < 130 \tag{D2}$$

$$G > R * 0.6 \tag{D3}$$

$$R > \text{skin red} + 30 \tag{D4}$$

$$G > R * 0.6 \tag{D5}$$

$$R <= \text{skin red} \tag{D6}$$

$$G > R * 0.7 \tag{D7}$$

$$(R <= \text{skin red} + 30) \wedge (R > \text{skin red}) \wedge (|G - B| > 10) \tag{D8}$$

$$\text{number of unprocessed pixels in image} = 0 \tag{D9}$$

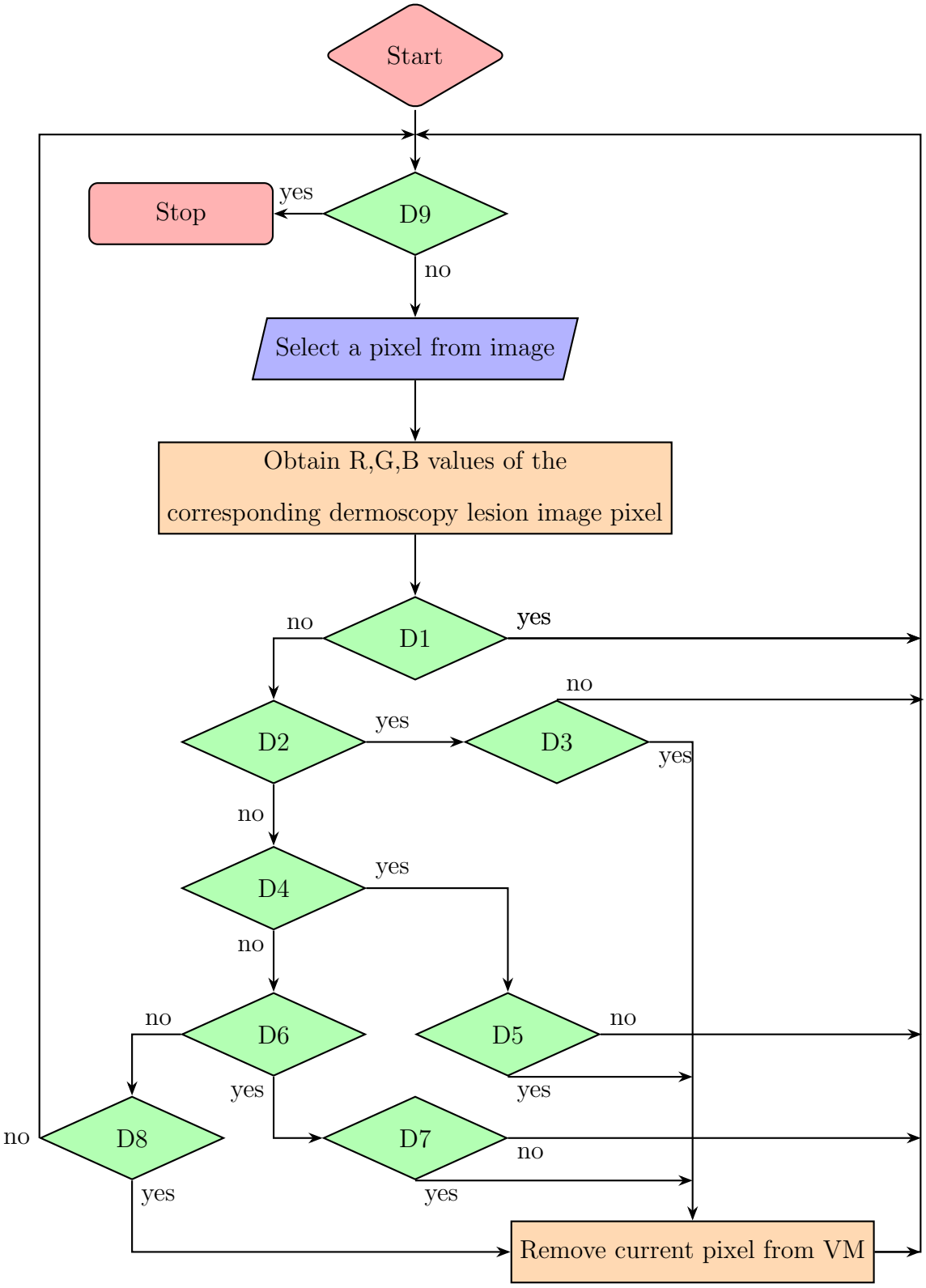
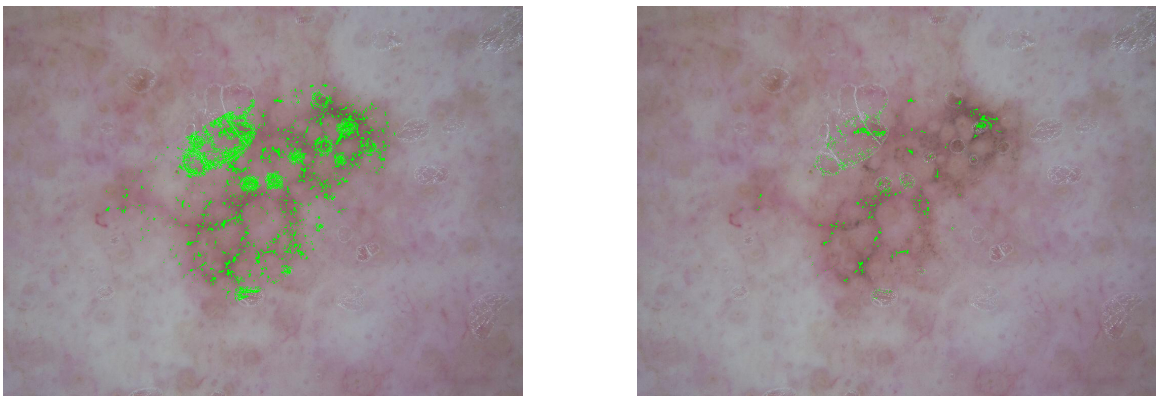


Figure 3.7. Flow diagram for section 2

3.6.3. Stage 3. In this stage, the processed vessel mask from Stage 1 and Stage 2 is taken as input. If $R < 1.1 * G \wedge R < 1.1 * B$, then the pixel should be eliminated from the vessel mask. An example of bubble filter is given in Figure 3.8.

$$R < 1.1 * G \wedge R < 1.1 * B \tag{3.6}$$



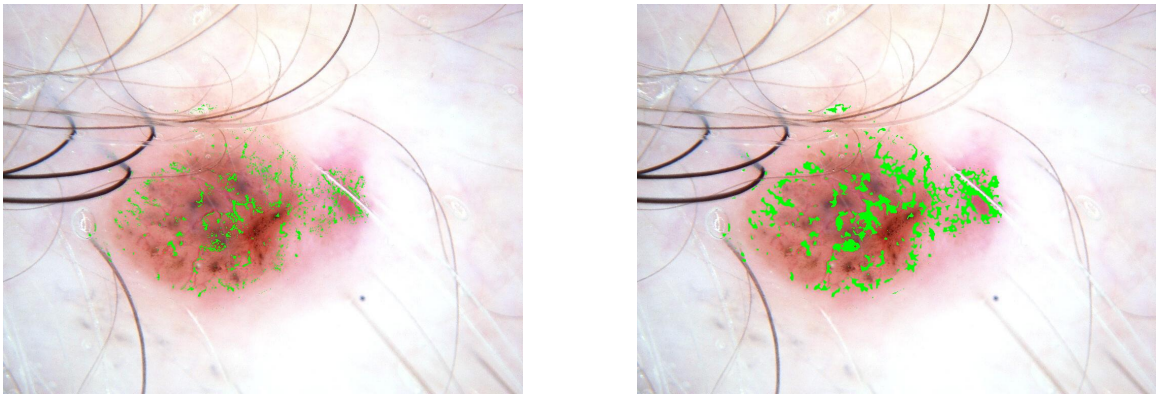
(a) Hair mask filtered image (b) Bubble mask filtered image

Figure 3.8. Vessel mask before and after bubble mask filter

3.7. VESSEL CONNECTOR

Vessels in the vessel mask may be broken because of the bubble filter or an imperfect hair mask. Dilating the vessel mask with a disk structure of radius 3 in Matlab followed by erosion with a disk structure of radius 2 was generally successful in connecting vessels [8]. Though it successfully connects vessels most of the time it also makes vessels larger than their original size and combines adjacent objects. If a vessel combines with noise and forms a large object, the blob filter may remove the

vessel. This stage also makes vessels wider than their original shape. One advantage of a broken vessel connector is its ability to fix the broken vessel problem caused by the color drop filter in some images. An example of vessel connector is shown in Figure 3.9.



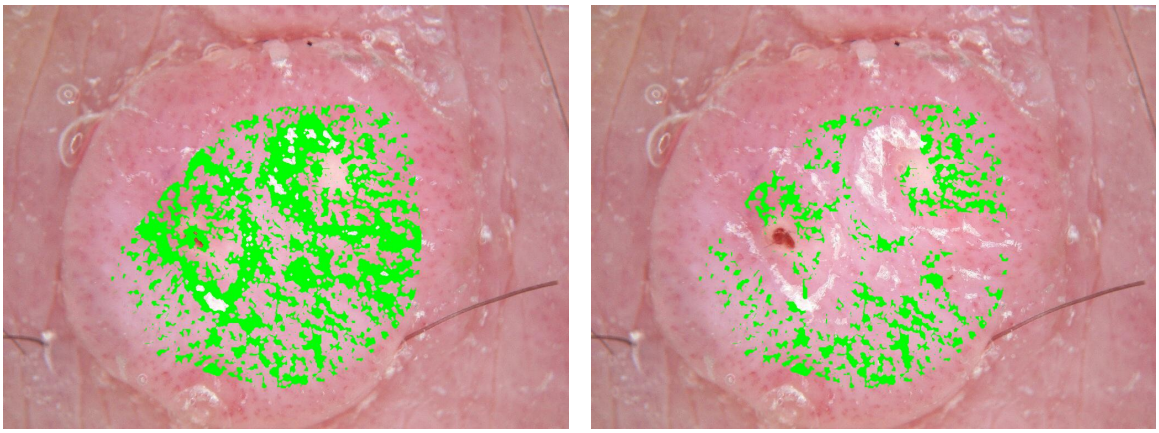
(a) Bubble mask filtered image (b) Vessel connector filtered image

Figure 3.9. Vessel mask before and after vessel connector filter

3.8. BLOB FILTER

A 37 x 37 all-ones mask is placed on every pixel of the vessel mask (VM). The area of the resulting mask is calculated, and if the resulting area is greater than 70% of the 37 x 37 unity mask area, then the entire connected object was removed from the vessel mask. An example of blob filter is shown in Figure 3.10.

$$Area = \sum_{i=-18}^{18} \sum_{j=-18}^{18} 1 * VM(x+i, y+j) \tag{3.7}$$



(a) Vessel connector mask filtered image (b) Blob filtered image

Figure 3.10. Vessel mask before and after blob filter

3.9. VESSEL BINDER

The vessel binder filter tries to connect broken vessels without much noise. A skeleton of vessel mask is created using bwmorph Matlab inbuilt command [8]. The obtained skeleton of the VM is dilated using a disk structural object of radius 3 using the imdilate command. A skeleton of the dilated image is generated and is added it back to the initial vessel mask. The final mask obtained from this step is then passed through a second color drop vessel detection filter. The vessel binder binding vessels is shown in Figure 3.11.

3.10. SECOND COLOR DROP VESSEL DETECTION FILTER

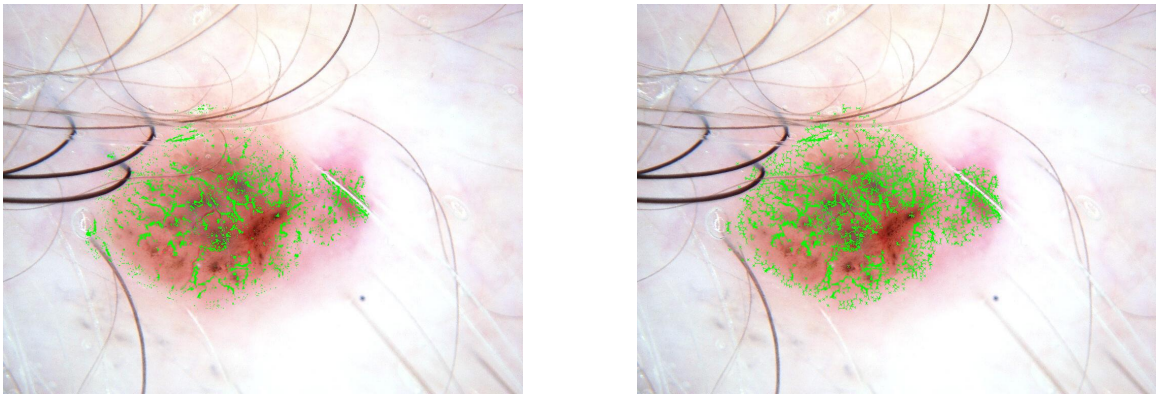
This filter uses the same code as the color drop vessel detection described before. This filter is used after the vessel binder. The vessel mask obtained from the vessel binder is dilated. The main difference for this algorithm is that instead of running the color drop vessel detection filter on every pixel inside the lesion of dermoscopy image, the color drop vessel detection runs on the dermoscopy image pixels which have a corresponding value of one in the dilated vessel mask. This filter

can regenerate vessels lost in previous steps but will also add noise. The performance of second color drop vessel detection filter is shown in Figure 3.12.

3.11. RANDOM NOISE FILTER

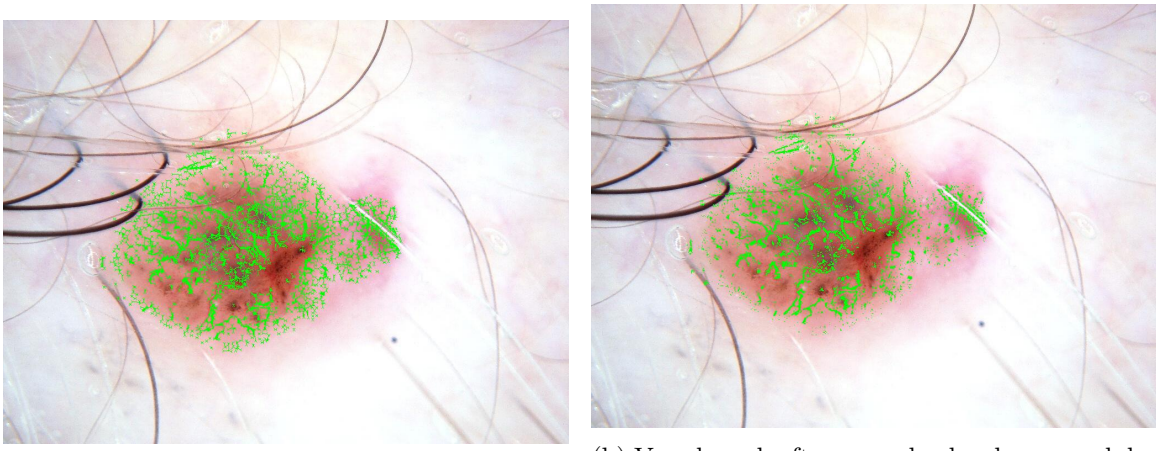
After all the above mentioned vessel detection and filtering noise is still left in the vessel mask. Each object in the vessel mask is detected using the BWLABEL Matlab inbuilt function. The detected object is skeletonized; the skeleton object area corresponds to its object length. If the length of the object is less than 30, then the object is removed from the vessel mask.

After filtering short noise in the VM by length, the remaining noise is filtered based on area. The vessel mask is passed through morphological closing with a disk structure of radius 1. Then, each object in the vessel mask is selected and its corresponding area is recorded. If the measured area is less than 40 pixels or greater than 20000, the object is removed. Performance of this filter can be seen in Figure 4.2h, Figure 5.2h, and Figure 6.2j.



(a) Vessel mask after hair filter (b) Vessel mask after vessel binder

Figure 3.11. Vessel mask before and after vessel binder



(a) Vessel mask after vessel binder

(b) Vessel mask after second color drop vessel detection filter

Figure 3.12. Vessel mask before and after second color drop vessel detection

4. CHENG METHOD

As described in the third chapter, different filters were used to generate the telangiectasia vessel mask. Cheng used filters in the order displayed in Figure 4.1 [6]. The color drop filter is found very effective in detecting most of the vessels. In some images the bubble filter used in the Cheng model was found to break vessels that are attached by using the broken vessel connector. But the Cheng Model relied on morphological functions which are insensitive to noise and vessels. These morphological functions may cause near by noise and a vessel to combine and form a single large object. Big objects may be removed by a blob filter which might result in loss of some important vessels. The visualization of vessel mask at each filter stage is shown in Figure 4.2.

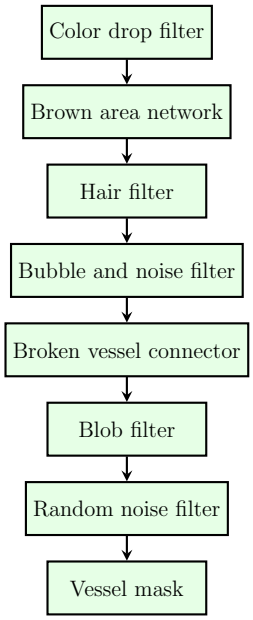
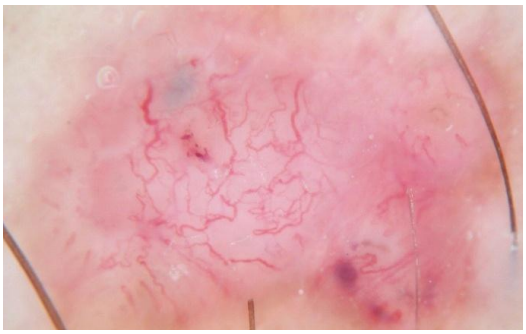
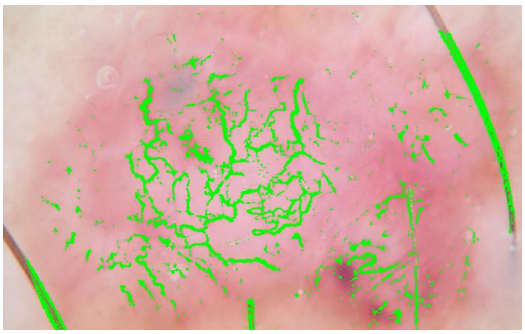


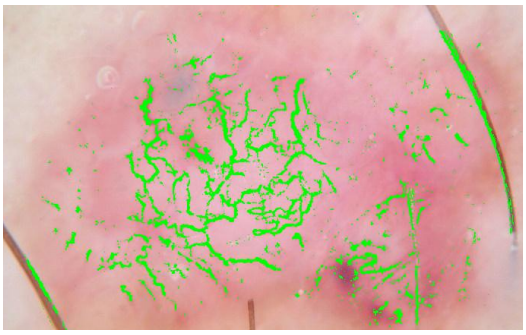
Figure 4.1. Cheng Model [6]



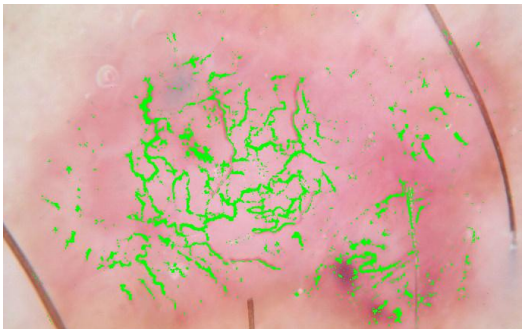
(a) Dermoscopy lesion image



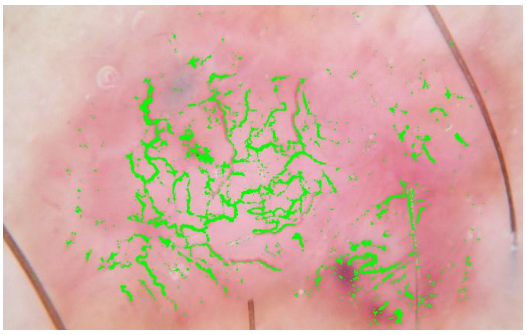
(b) Vessel mask after colordrop filter



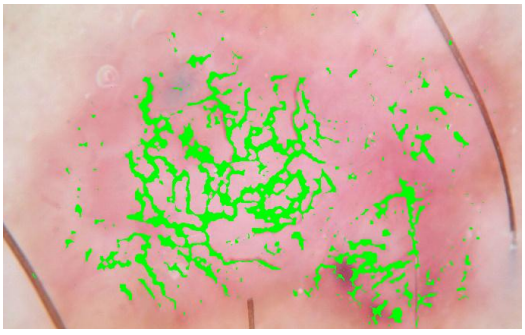
(c) Vessel mask after brown area network filter



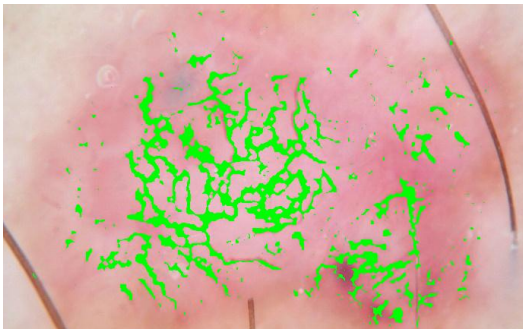
(d) Vessel mask after hair mask filter



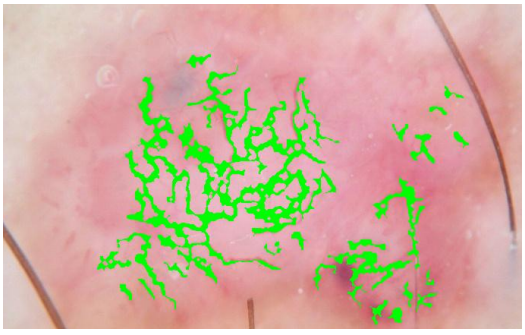
(e) Vessel mask after bubble filter



(f) Vessel mask after broken vessel connect filter



(g) Vessel mask after blob filter



(h) Vessel mask after short noise filter

Figure 4.2. Cheng Method step.by.step visualization

5. MODIFIED MODEL 1

The motivation for Modified Model 1 (MM1) was to let the vessels maintain their original shape and have less noise in the final vessel mask. This resulted in the removal of bubble filter and the broken vessel connect filters used by Cheng. Two other filters the red chromaticity filter and the white area filters were introduced to remove noise. The final vessel mask has less noise compared to the Cheng Model. If the color drop filter generates a segmented vessel instead of a continuous vessel, some part of the vessels may be lost due to the random noise filter. The order of filters resulting in MM1 is shown in Figure 5.1 where as vessel mask at each filter is proved in Figure 5.2.

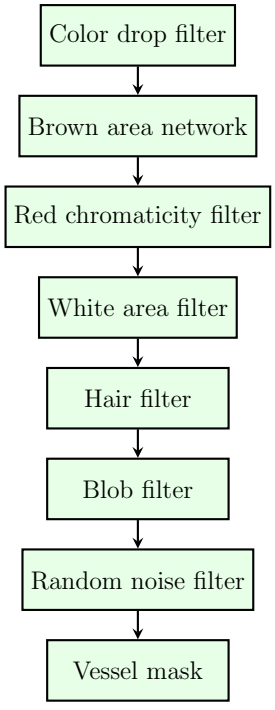
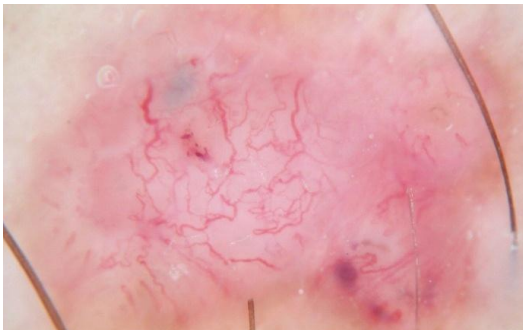
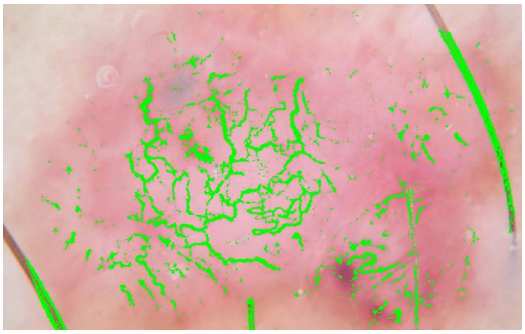


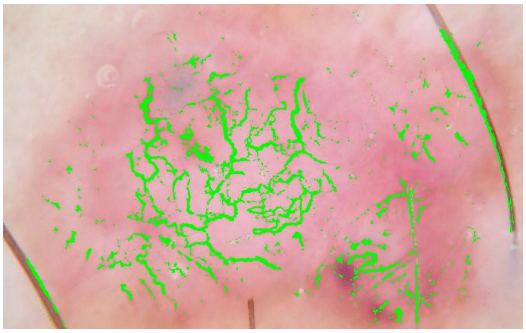
Figure 5.1. Modified Model 1



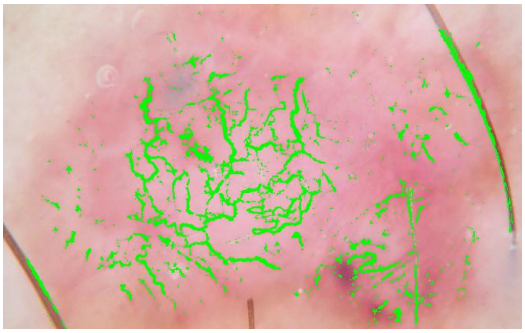
(a) Dermoscopy lesion image



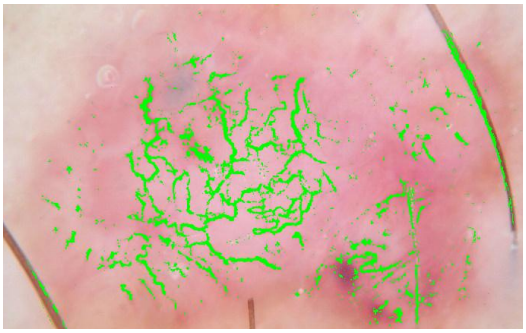
(b) Vessel mask after colordrop filter



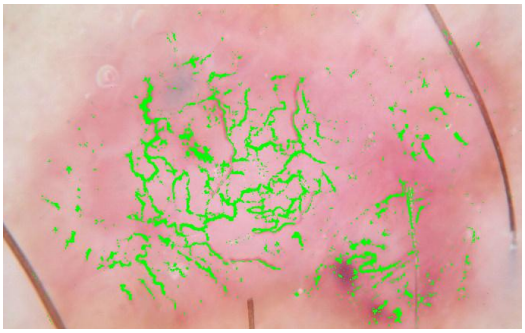
(c) Vessel mask after brown area network filter



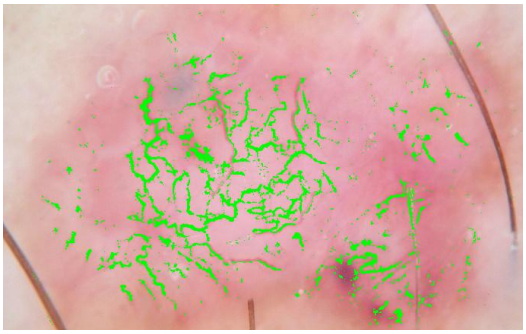
(d) Vessel mask after red chromaticity filter



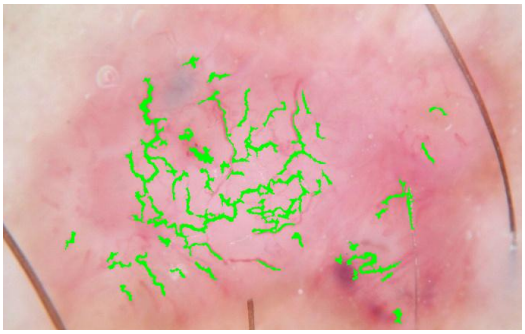
(e) Vessel mask after white area filter



(f) Vessel mask after hair mask filter



(g) Vessel mask after blob filter



(h) Vessel mask after short noise filter

Figure 5.2. Modified Model 1 step.by.step visualization

6. MODIFIED MODEL 2

As discussed in previous sections, for some images the color drop filter may produce segmented vessels which are removed by the random noise filter and the hair mask. Modified Model 2 use vessel binder and second color drop stages to regenerate some of the lost vessels which can be observed from Figure 3.11a and Figure 3.12b. Noise found in this model is more than in Modified Model 1 and less than in the Cheng Model. The order of filters resulting in MM1 is shown in Figure 6.1 where as vessel mask at each filter is proved in Figure 6.2.

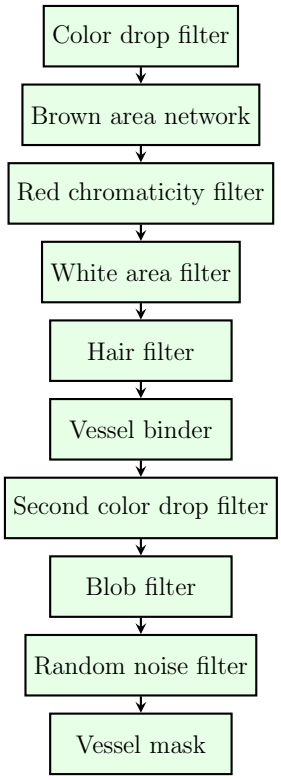
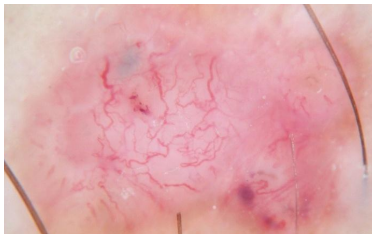
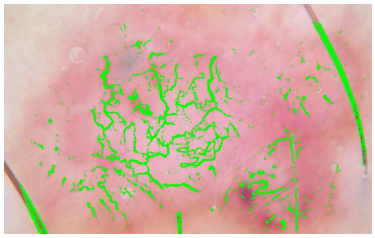


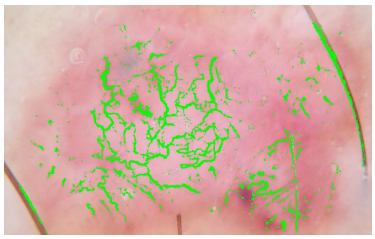
Figure 6.1. Modified Model 2



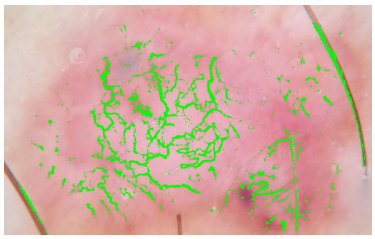
(a) Dermoscopy lesion image



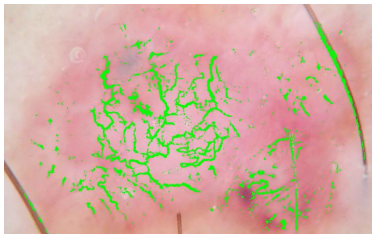
(b) Vessel mask after colordrop filter



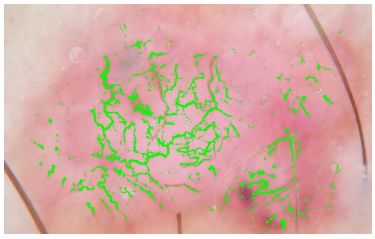
(c) Vessel mask after brown area network filter



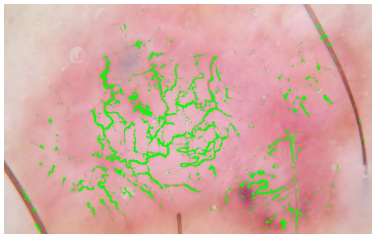
(d) Vessel mask after red chromaticity filter



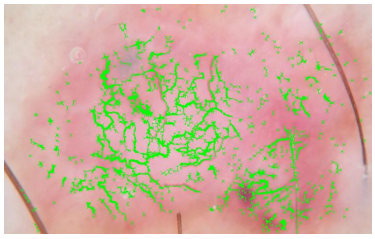
(e) Vessel mask after white area filter



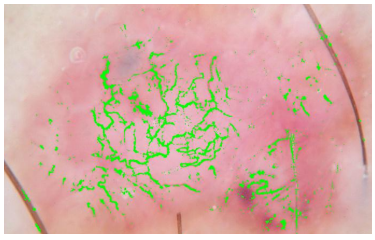
(f) Vessel mask after hair mask filter



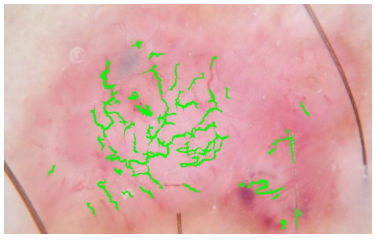
(g) Vessel mask after blob filter



(h) Vessel mask after vessel binder



(i) Vessel mask after second color drop filter



(j) Vessel mask after short noise filter

Figure 6.2. Modified Model 2 step.by.step visualization

7. FEATURE EXTRACTION

Features are generated using objects in the vessel mask. Extracted features were used for the classification of BCC. Two different types of the features sets were developed. One is based on the Cheng paper [6]. Extra features like skin red, skin green, skin blue and lesion areas were added to the Cheng features and appear in the Table 7.1 inaddition to initial Cheng features. The other set of features is shown in Table 7.2 and consists of 83 features which include some of the features from Table 7.1.

Table 7.1. Features used by Cheng

Index	**Feature name**	**Explanation**
1	Number of objects	Number of vessels that exist in the final vessel mask
2	Maximum length	Maximum length of all the vessels in the vessel mask
3	Maximum Area	Area of the vessel with largest area in the vessel mask
4	Average Length	Average length of all vessels in the vessel mask
5	Average area	Average area of all vessel areas in vessel mask
6	Average width	Average width of all vessels
7	Std width	Standard deviation of the width of all vessels

Table 7.1. Features used by Cheng (cont.)

Index	**Feature name**	**Explanation**
8	Minimum eccentricity	Minimum eccentricity of all vessels
9	Maximum eccentricity	Maximum eccentricity of all vessels
10	Average eccentricity	Average eccentricity of all vessels
11	Lesion area	Lesion area of the basal cell carcinoma image
12 to 21	Eroded objects	Number of object present after the vessel mask eroded with a disk structure of radius 1 to 10
22 to 31	Eroded objects area	Area of objects present after the vessel mask eroded with a disk structure of radius 1 to 10
32	Skin red	Average of the red values outside of lesion
33	Skin green	Average of the green values outside of lesion
34	Skin blue	Average of the blue values outside of lesion

7.1. GENERATION OF OBJECT SURROUND

Every object from the vessel mask is extracted and dilated with a disk structuring element of radius 12 (d1) and radius 5 (d2) resulting two dilation variants of the object. Removing d2 from d1 gives the object surround. The object surround may contain vessel pixels so they are eliminated by using the vessel mask.

7.2. OBJECT SURROUND FEATURES

Regions surrounding the vessels can also be useful in differentiating basal cell carcinoma; hence Features 1 to 24 mentioned in Table 7.2 were also calculated for the surround area of the vessel and become Features 25 to 48 in Table 7.2. Similarly, way Features 51 to 54 were applied on the surrounds of the objects resulting in Features 55 to 58.

Table 7.2. Extended feature set

Index	**Feature name**	**Explanation**
1	Maximum red of objects	Maximum red of every vessel is extracted into an array. Average of the obtained array is used as the maximum red feature
2	Maximum green of objects	Maximum green of every vessel is extracted into an array. Average of the obtained array is used as the maximum green feature
3	Maximum blue of objects	Maximum blue of every vessel is extracted into an array. Average of the obtained array is used as the maximum blue feature
4	Minimum red of objects	Minimum red of every vessel is extracted into an array. Average of the obtained array is used as the minimum red feature
5	Minimum green of objects	Minimum green of every vessel is extracted into an array. Average of the obtained array is used as the minimum green feature

Table 7.2. Extended feature set (cont.)

Index	**Feature name**	**Explanation**
6	Minimum blue of objects	Minimum blue of every vessel is extracted into an array. Average of the obtained array is used as the minimum blue feature
7	Average red of objects	Red value of every pixel in a single vessel of vessel mask is added and divided by the total number of pixels in the vessel to obtain average red of a vessel. Average red of all the vessels is calculated similarly and accumulated into an array. Average of the array is performed to obtain the desired feature
8	Average green of objects	Green value of every pixel in a single vessel of vessel mask is added and divided by the total number of pixels in the vessel to obtain average green of a vessel. Average green of all the vessels is calculated similarly and accumulated into an array. Average of the array is performed to obtain the desired feature
9	Average blue of objects	Blue value of every pixel in a single vessel of vessel mask is added and divided by the total number of pixels in the vessel to obtain average blue of a vessel. Average blue of all the vessels is calculated similarly and accumulated into an array. Average of the array is performed to obtain the desired feature

Table 7.2. Extended feature set (cont.)

Index	**Feature name**	**Explanation**
10	Standard deviation of red for objects	Standard deviation of red for every single vessel is calculated. Mean of all the standard deviation corresponding to every vessel gives the desired feature
11	Standard deviation of green for objects	Standard deviation of green for every single vessel is calculated. Mean of all the standard deviation corresponding to every vessel gives the desired feature
12	Standard deviation of blue for objects	Standard deviation of blue for every single vessel is calculated. Mean of all the standard deviation corresponding to every vessel gives the desired feature
13 to 24	same as 1 to 12	All the above mentioned methods from 1 to 12 applied in HSV plane to obtain Features 13 to 24
25 to 48	Same as 1 to 24 for surround of objects	All the methods from 1 to 24 are performed on the surround of objects in the vessels to obtain desired Features 25 to 48
49	Area	Mean of area of all the vessels in vessel mask
50	Average width of the vessel	Mean of average width of all the vessels in vessel mask
51	Average eccentricity	Mean of average eccentricity of all the vessels in vessel mask

Table 7.2. Extended feature set (cont.)

Index	**Feature name**	**Explanation**
52	Minimum eccentricity	Mean of minimum eccentricity of all the vessels in vessel mask
53	Maximum eccentricity	Mean of maximum eccentricity of all the vessels in vessel mask
54	Standard deviation of eccentricity	Mean of standard deviation of eccentricity of all the vessels in vessel mask
55 to 58	Same as 51 to 54 for surround objects	Methods used for 51 to 54 are used for the surrounding objects in the vessel mask
59 to 68	Eroded objects	Number of object present after the vessel mask eroded with a disk structure of radius 1 to 10
69 to 78	Eroded objects area	Total area of objects present after the vessel mask eroded with a disk structure of radius 1 to 10
79	Number of objects	Number of vessels that exist in the final vessel mask
80	Lesion area	Lesion area of the dermoscopy lesion image
81	Skin red	Average of the red values outside of lesion
82	Skin green	Average of the green values outside of lesion
83	Skin blue	Average of the blue values outside of lesion

8. BCC CLASSIFICATION

Detection of basal cell carcinoma is a binary classification problem. Logistic regression is a widely used, simple, and reliable tool for binary classification; hence it is employed for analysis.

8.1. LOGISTIC REGRESSION

Dataset $\boldsymbol{X}$ is a matrix with dimensions d and N samples such that the d dimension also included a constant. let Y be a binary vector with label 1 if y_i belong to a positive set or label 0 if y_i belong to a negative set. The risk score of the data set is given by Equation 8.1, where $\boldsymbol{W}$ gives the weights for $\boldsymbol{X}$. The logistic regression function is defined by Equation 8.2:

$$z = \mathbf{W}^T\mathbf{X} \tag{8.1}$$

$$\psi(z) = \frac{1}{1 + e^{-}z} \tag{8.2}$$

Based on the likelihood of X and Y, minimizing the in-sample error of training data given by equation 8.3 gives the desired hypothesis [9].

$$E_{in} = \frac{1}{N}\sum_{n=1}^{N} \log\left(1 + e^{-y_n \mathbf{W}^T \mathbf{X}_n}\right) \tag{8.3}$$

Minimization of equation 8.3 is done using iterative gradient descent approach as mentioned by Abu-Mostafa [9]. The algorithm for the minimization is given below [9] using gradient descent.

Algorithm 1 Finding optimal W

1: Initialize weights at i=0 to W(0)
2: **for** i=0,1,2,3..... **do**
3: Compute the gradient

$$\bigtriangledown E_{in} = \frac{-1}{N} \sum_{n=1}^{N} \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{W}^T(i) \mathbf{x}_n}}$$

4: Update the weights : $\mathbf{W}$(i+1)= $\mathbf{W}$(i)-$\eta \bigtriangledown E_{in}$
5: Iterate through the for loop till stopping criteria are met
6: Get the final weights $\mathbf{W}$

8.2. SAS IMPLEMENTATION

Stepwise logistic regression with one fold cross validation is implemented in SAS software version 9.4. The builtin functions PROC and MODEL are used to construct the model.

8.3. STEP-WISE FEATURE SELECTION

Stepwise selection is used for feature selection. Stepwise selection starts with no features in the model. The feature with the smallest P-value that is less than SLENTRY enters the model. Next another feature with a p-value smaller than SLENTRY enters the model. Once the second feature enters the model P-values of any feature greater than the specified SLSTAY exit the model. Selection stops when none of the features outside the model have a p-value less than SLENTRY and all the features in the model have a p-value less than SLSTAY [10] [11] [12].

8.4. DATA DISTRIBUTION FOR BCC CLASSIFICATION

A total of 1024 images were used in the analysis; 527 NonBCC images and 229 BCC images were randomly selected for training while the remaining 193 NonBCC images and 75 BCC were used for testing as shown in below tables. A random seed generator is used to keep the random distribution constant for all logistic regression models. Data distributions are shown in the Table 8.1 and Table 8.2.

Table 8.1. Training set distribution

Type	Frequency	percentage
NonBCC	537	69.71
BCC	229	30.27

Table 8.2. Test set distribution

Type	Frequency	percentage
NonBCC	193	72.01
BCC	75	27.99

8.5. CLASSIFICATION RESULTS

In the Feature Extraction section, Table 7.1 and Table 7.2 display two feature sets. Development of those features happened gradually. Initially only 14 features

were used for basal cell carcinoma classification, those features increased to 34 features, 63 features, and finally to 83 features. Details of the features are shown in Table 8.3. Cheng initially used 10 features later upgraded to 30 features as shown in her paper [6].

Table 8.3. Different features used for BCC Classification

Features	Explanation
14	All features from Table 7.1 excluding 12 to 31
34	All features from Table 7.1
63	All features from Table 7.2 excluding 59 to 78
83	All features from Table 7.2

Three models, the Cheng Model, Modified Model 1 and Modified Model 2 were proposed for the generation of the vessel mask. Cheng has used the Cheng Model with 10 and 30 features mentioned in Table 8.4. The four feature sets described in Table 8.3 were generated for each model. These features were used for training and testing using stepwise logistic regression. Apart from the features, two variations of SLENTRY and SLSTAY were also used. All this combinations resulted in 48 models for BCC classification.

Table 8.4. Features used by Cheng

Features	Explanation
10	1 to 10 features from Table 7.1
30	All features from Table 7.1 except Features 11, 32, 33, and 34

Training and test ROC curves for the features 14, 34, 63, and 83 mentioned in Table 8.3 with SLENTRY (SLE) = 0.01 and SLSTAY (SLS)= 0.15 are shown in Figure 8.1. Note that in some graphs there appear to be only three curves because two of the curves overlap completely. It can be observed that there is an increase in BCC classification performance with an increase in number of features while their is little to no difference in ROC curves for 63 and 83 features in all six plot shown in Figure 8.1. From the test ROC curves, it can be observe that all models were generalized. MM1 tends to perform better than the other models with 14 and 34 features in terms of AUC of the training model. The AUC of the CM training model is 0.873 for both 63 and 83 features. The AUC of the MM2 training model is almost as high at 0.87021 for 63 featues and 0.87093 for 83 features.

Training and test ROC curves for the features 14, 34, 63, and 83 features mentioned in Table 8.3 with SLE = 0.3 and SLS= 0.3 are shown in Figure 8.2. As the SLE and SLS values increased, more features were selected for the final model. It can be observed that there is an increase in BCC classification performance with increase in number of features while there is little to no difference in ROC curves for 63 and 83 features in all six plots shown in Figure 8.2. It can be observed from the ROC test curves that all models were generalized. MM1 tends to perform better than the other models with 14 and 34 features in terms of AUC of the training model. MM2 with the 83 features dataset had the highest training AUC among all 24 BCC classification training models shown in Figure 8.1 and Figure 8.2.

Training and test ROC curves for the features 14, 34, 63, and 83 mentioned in Table 8.3 with up to two variable interactions between features and with SLE = 0.01 and SLS= 0.15 are shown in Figure 8.3. As the SLE and SLS values were small, fewer features will be in the final model. Products between all possible combinations of two features were also considered as features in addition to the original feature set. This is the major difference for classification models in Figure 8.3 compared to Figure 8.1.

It can be observed that there is an increase in BCC classification performance with an increase in number of features while their is little to no difference in ROC curves for 63 and 83 features in all six plots shown in Figure 8.3. From the test ROC curves, it can be observed that all models were generalized. It can aslo be observed that there is no significant improvement from previous models.

Training and test ROC curves for the features 14, 34, 63, and 83 mentioned in Table 8.3 with up to two variable interactions between features and with SLE = 0.3 and SLS= 0.3 are shown in Figure 8.4. Products between all possible combinations of two features were also considered as features in addition to the original feature set. This is the major difference for classification models in Figure 8.2 compared to Figure 8.4. It can be observed that there is an increase in BCC classification performance with an increase in number of features while their is little to no difference in ROC curves for 63 and 83 features in all six plots shown in Figure 8.4. Their was no significant improvement from previous models. Though classification models in Figure 8.4e show higher AUC, they are clearly overfitted, as evident in Figure 8.4f.

8.6. TEST RESULTS AT A SPECIFIC CUTOFF

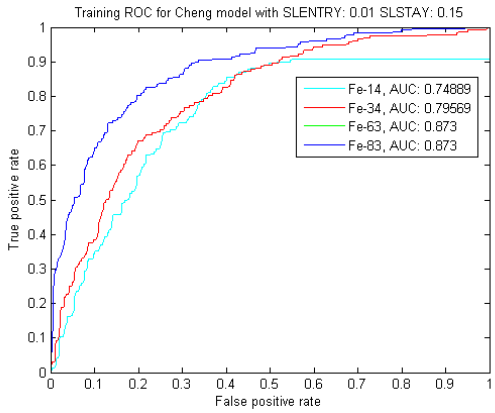
Logistic regression has an output from 0 to 1. A cutoff of 0.04 has been selected which poses 98% sensitivity for most of the models. Test results for all the classification models are provided in Table 8.6. Most of the test results maintained sensitivity near 98%, while some have deviated significantly because of overfitting.

8.7. CHENG OLD RESULTS

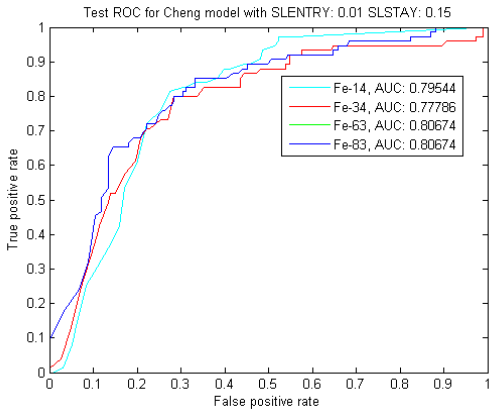
BCC classification with 30 features proposed by Cheng [6] was performed using the Cheng model by stepwise logistic regression with SLE and SLS at 0.3. Two ROC curves, one without interaction and the other with interactions are shown in Figure 8.5, while the test performance is shown in Table 8.5.

Table 8.5. Cheng 30 features test set

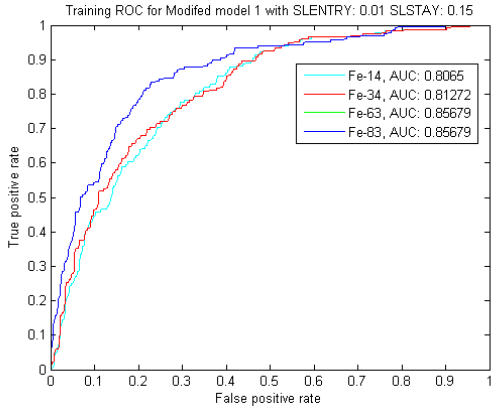
obs	FEA	INT	SLE	SLS	SEN	SPC	ACC	fscore
1	30	No	0.3	0.3	100	0	27.985	43.7317
2	30	Yes	0.3	0.3	92	10.3626	33.20895	43.533



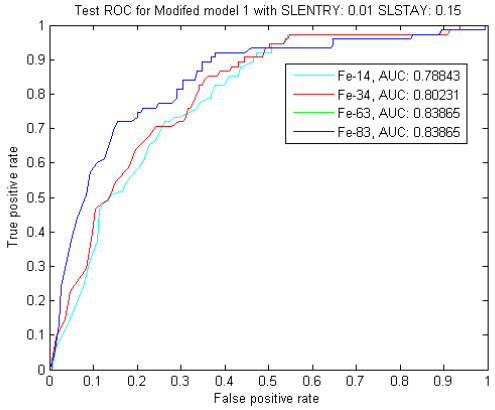
(a) Cheng Model Training ROC



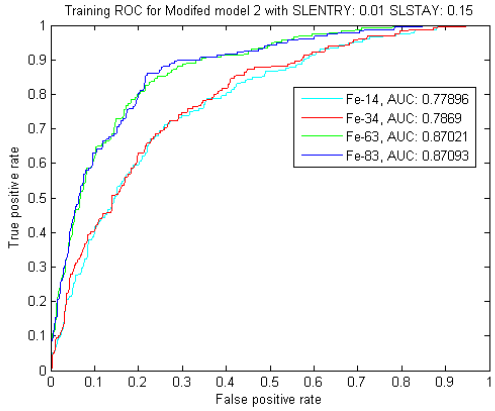
(b) Cheng Model Test ROC



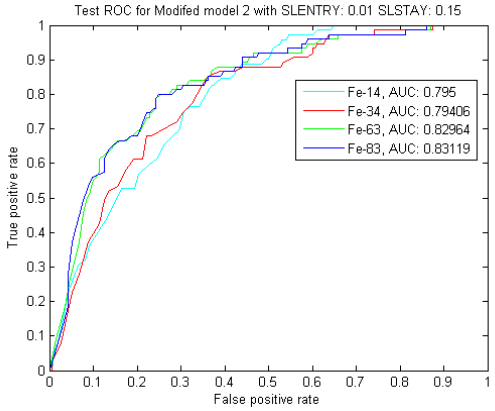
(c) MM1 Training ROC



(d) MM1 Test ROC

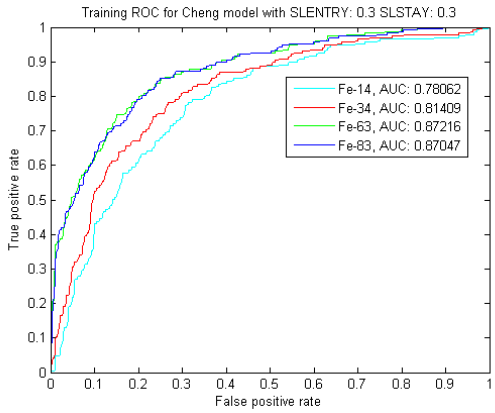


(e) MM2 Training ROC

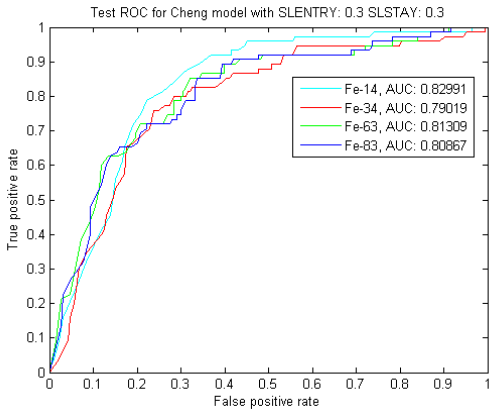


(f) MM2 Test ROC

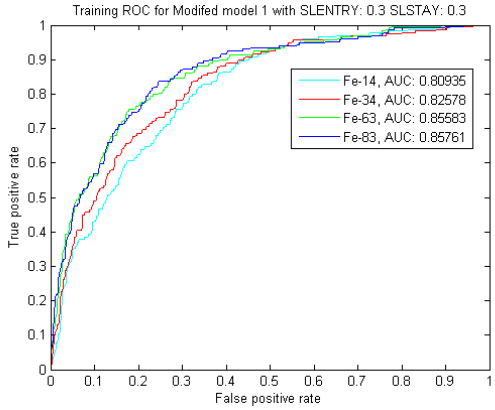
Figure 8.1. ROC curves for the Cheng model, MM1, and MM2 with SLENTRY=0.01 and SLSTAY=0.15



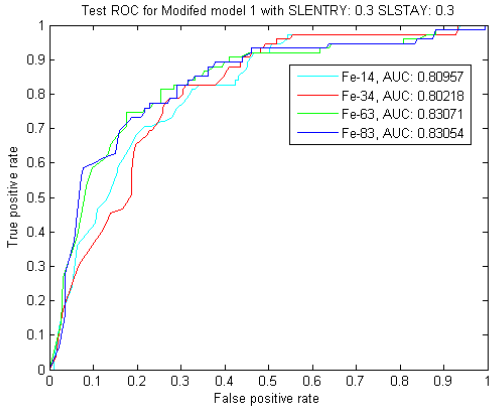
(a) Cheng Model Training ROC



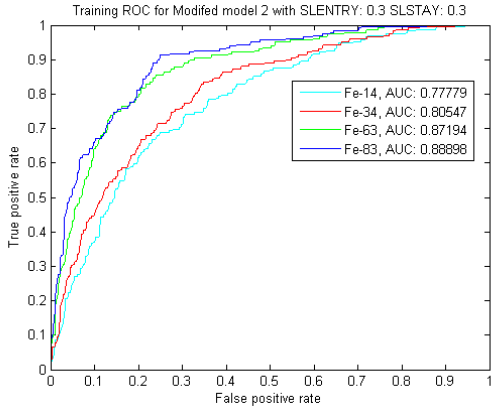
(b) Cheng Model Test ROC



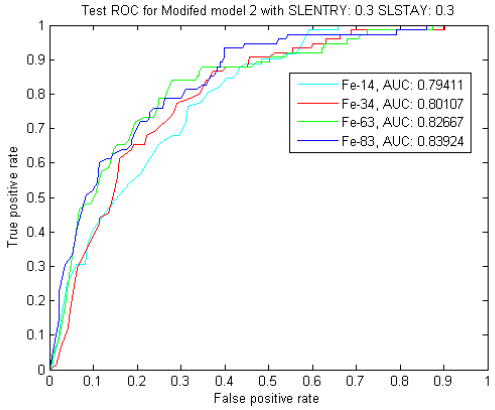
(c) MM1 Training ROC



(d) MM1 Test ROC

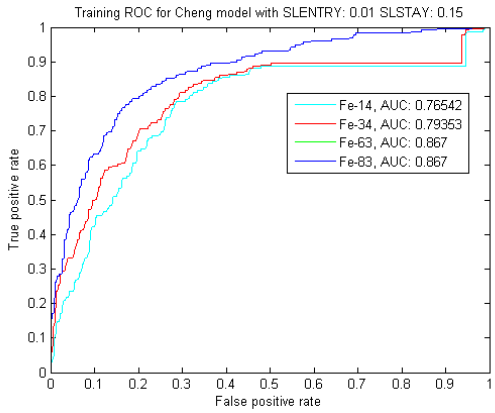


(e) MM2 Training ROC

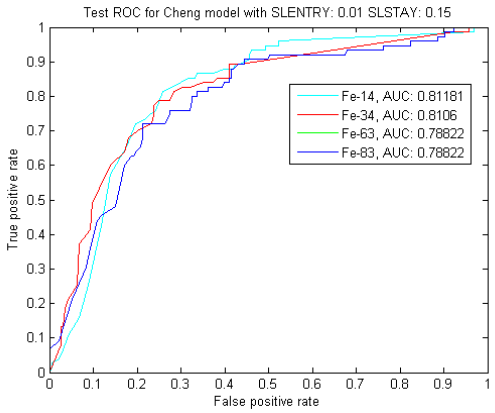


(f) MM2 Test ROC

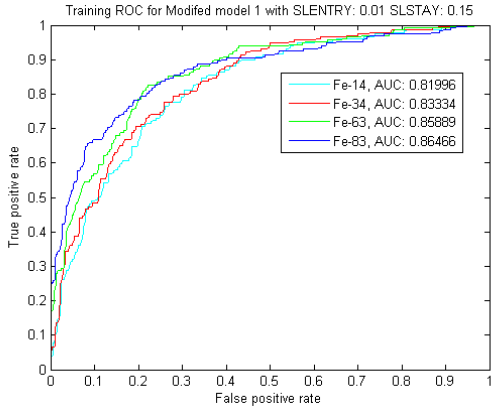
Figure 8.2. ROC curves for the Cheng model, MM1 and MM2 with SLENTRY=0.3 and SLSTAY=0.3



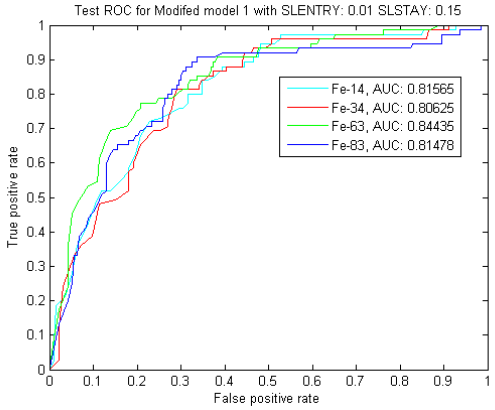
(a) Cheng Model Training ROC



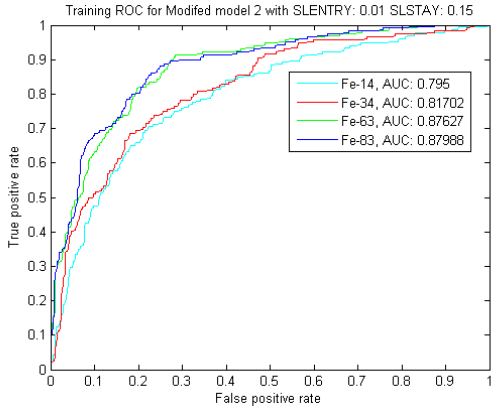
(b) Cheng model Test ROC



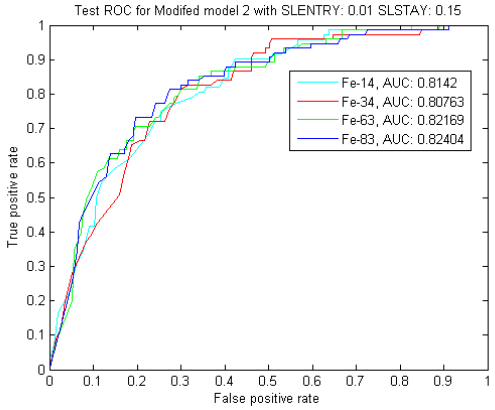
(c) MM1 Training ROC



(d) MM1 Test ROC

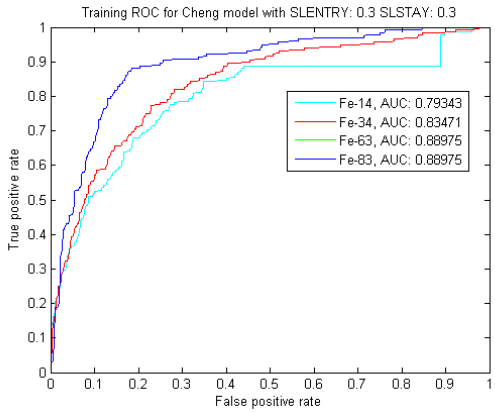


(e) MM2 Training ROC

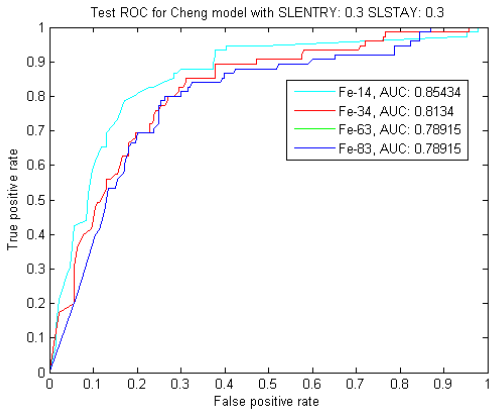


(f) MM2 Test ROC

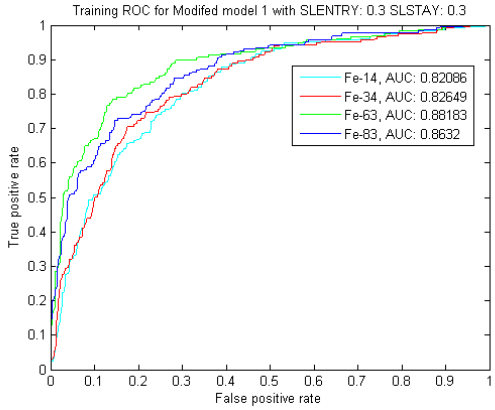
Figure 8.3. ROC curves for the Cheng model, MM1 and MM2 with SLENTRY=0.01 and SLSTAY=0.15 and up to two variable interaction between features



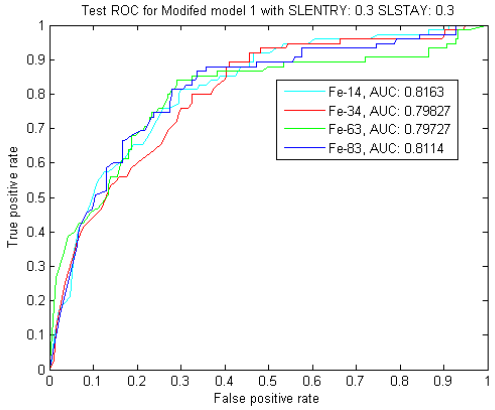
(a) Cheng Model Training ROC



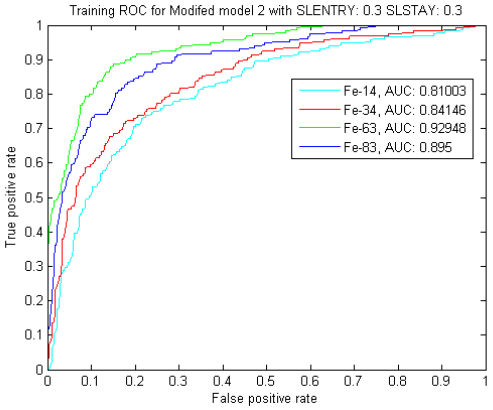
(b) Cheng Model Test ROC



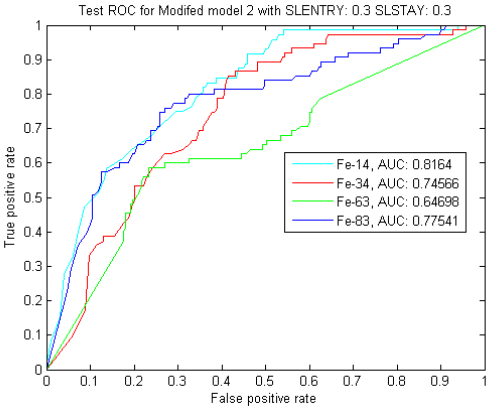
(c) MM1 Training ROC



(d) MM1 Test ROC



(e) MM2 Training ROC



(f) MM2 Test ROC

Figure 8.4. ROC curves for Cheng model, MM1 and MM2 with SLENTRY=0.3 and SLSTAY=0.3 and up to two variable interaction between features

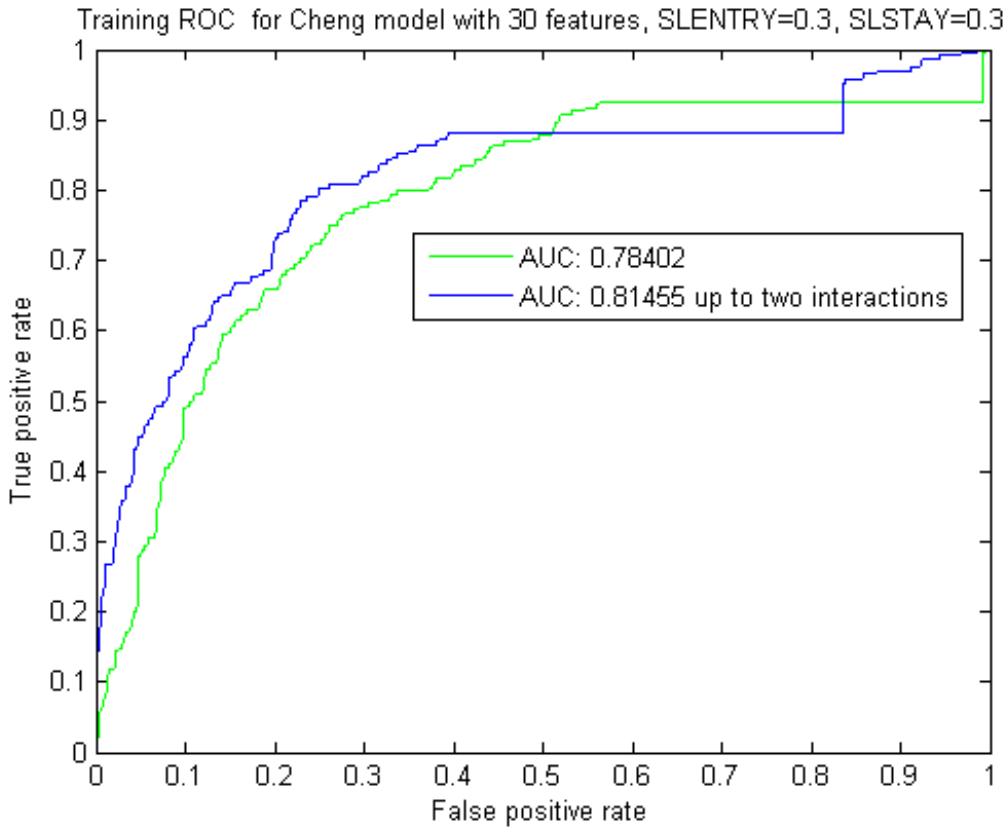


Figure 8.5. Cheng training results with 30 features

Table 8.6. Test set results for all BCC models

obs	**INT**	**FEA**	**SLE**	**SLS**	**SEN**	**SPC**	**ACC**	**fscore**
1	No	14	0.01	0.15	100	0	27.9851	43.7318
2	No	34	0.01	0.15	100	0	27.9851	43.7318
3	No	63	0.01	0.15	98.6667	2.5907	29.4776	43.9169
4	No	83	0.01	0.15	98.6667	2.5907	29.4776	43.9169
5	No	14	0.01	0.15	100	0	27.9851	43.7318
6	No	34	0.01	0.15	100	0	27.9851	43.7318

Table 8.6. Test set results for all BCC models (cont.)

7	No	63	0.01	0.15	100	11.9171	36.5672	46.875
8	No	83	0.01	0.15	100	11.9171	36.5672	46.875
9	No	14	0.01	0.15	100	0.5102	27.2388	42.4779
10	No	34	0.01	0.15	100	1.0363	28.7313	43.9883
11	No	63	0.01	0.15	98.6667	12.9534	36.9403	46.6877
12	No	83	0.01	0.15	98.6667	13.9896	37.6866	46.9841
13	No	14	0.3	0.3	100	0	27.9851	43.7318
14	No	34	0.3	0.3	100	1.0363	28.7313	43.9883
15	No	63	0.3	0.3	98.6667	4.1451	30.597	44.3114
16	No	83	0.3	0.3	98.6667	4.1451	30.597	44.3114
17	No	14	0.3	0.3	100	0	27.9851	43.7318
18	No	34	0.3	0.3	100	0	27.9851	43.7318
19	No	63	0.3	0.3	98.6667	12.4352	36.5672	46.5409
20	No	83	0.3	0.3	98.6667	12.4352	36.5672	46.5409
21	No	14	0.3	0.3	100	0.5102	27.2388	42.4779
22	No	34	0.3	0.3	100	2.5907	29.8507	44.3787
23	No	63	0.3	0.3	98.6667	16.5803	39.5522	47.7419
24	No	83	0.3	0.3	98.6667	17.6166	40.2985	48.0519
25	Yes	14	0.01	0.15	100	0.51813	28.3582	43.8596
26	Yes	34	0.01	0.15	100	1.0363	28.7313	43.9883
27	Yes	63	0.01	0.15	100	3.6269	30.597	44.6429
28	Yes	83	0.01	0.15	97.3333	9.3264	33.9552	45.2012
29	Yes	14	0.01	0.15	100	1.5544	29.1045	44.1176
30	Yes	34	0.01	0.15	98.6667	5.6995	31.7164	44.713
31	Yes	63	0.01	0.15	97.3333	11.399	35.4478	45.768

Table 8.6. Test set results for all BCC models (cont.)

32	Yes	83	0.01	0.15	97.3333	11.399	35.4478	45.768
33	Yes	14	0.01	0.15	100	1.0204	27.6119	42.6036
34	Yes	34	0.01	0.15	100	5.1813	31.7164	45.045
35	Yes	63	0.01	0.15	98.6667	14.5078	38.0597	47.1338
36	Yes	83	0.01	0.15	98.6667	16.5803	39.5522	47.7419
37	Yes	14	0.3	0.3	100	1.5544	29.1045	44.1176
38	Yes	34	0.3	0.3	100	2.0725	29.4776	44.2478
39	Yes	63	0.3	0.3	93.3333	12.4352	35.0746	44.586
40	Yes	83	0.3	0.3	98.6667	7.2539	32.8358	45.122
41	Yes	14	0.3	0.3	100	1.5544	29.1045	44.1176
42	Yes	34	0.3	0.3	98.6667	6.7358	32.4627	44.9848
43	Yes	63	0.3	0.3	94.6667	18.1347	39.5522	46.7105
44	Yes	83	0.3	0.3	94.6667	18.1347	39.5522	46.7105
45	Yes	14	0.3	0.3	100	4.0816	29.8507	43.3735
46	Yes	34	0.3	0.3	97.3333	18.1347	40.2985	47.7124
47	Yes	63	0.3	0.3	61.3333	58.5492	59.3284	45.7711
48	Yes	83	0.3	0.3	92	24.8705	43.6567	47.7509

9. CONCLUSION

It can be observed from Figure 8.1, Figure 8.2, Figure 8.3 and Figure 8.4 that MM1 has performed better with fewer features (14, 34) while MM2 and the Cheng Model performed better with more features. From all the explored 48 classification models Figure 8.2e has relatively better performance in the training set and test set. MM2 tends to have less noise than CM and more noise than MM1. In some cases MM2 preserved most of telangiectasia vessels better than MM1 and the Cheng model. This implies that detection of vessels and reduction of noise in the final vessel mask has boosted the performance of detecting BCC along with an increase in the number of features.

10. FUTURE WORK

1. Though sensitivity above 98% is achieved, methods to improve specificity must be explored.

2. Filters mentioned in the third chapter had static thresholds. As lesion are dynamic in nature, having a dynamic threshold can reduce noise in the vessel mask.

3. The performance of the model needs to be analyzed using other classification techniques such as lasso regression, neural networks and support vector machine SVM.

4. The performance of the model should be observed based on other feature selection methods like PCA.

5. A new hair mask should be developed as the present hair mask was found to remove some of the main telangiectasia in the vessel mask.

6. More experimentation needs to be done with the green and blue plane drops for the Second color drop vessel detection filter so that more vessels are detected even at the cost of extra noise.

BIBLIOGRAPHY

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA: a cancer journal for clinicians*, vol. 66, no. 1, pp. 7–30, 2016.

[2] A. C. Society, "Cancer facts & figures 2016," *American Cancer Society. Atlanta*, 2016.

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

[4] C. Wong, R. Strange, and J. Lear, "Basal cell carcinoma," *British Medical Journal*, vol. 327, no. 7418, p. 794, 2003.

[5] T. S. Housman, S. R. Feldman, P. M. Williford, A. B. Fleischer, N. D. Goldman, J. M. Acostamadiedo, and G. J. Chen, "Skin cancer is among the most costly of all cancers to treat for the medicare population," *Journal of the American Academy of Dermatology*, vol. 48, no. 3, pp. 425–429, 2003.

[6] B. Cheng, D. Erdos, R. J. Stanley, W. V. Stoecker, D. A. Calcara, and D. D. Gómez, "Automatic detection of basal cell carcinoma using telangiectasia analysis in dermoscopy skin lesion images," *Skin Research and Technology*, vol. 17, no. 3, pp. 278–287, 2011.

[7] 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.

[8] M. U. Guide, "The mathworks," *Inc., Natick, MA*, vol. 5, p. 333, 1998.

[9] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning from data*, vol. 4. AMLBook Singapore, 2012.

[10] R. R. Hocking, "A biometrics invited paper. the analysis and selection of variables in linear regression," *Biometrics*, vol. 32, no. 1, pp. 1–49, 1976.

[11] S. I. Inc, *SAS user's guide statistics, version 5 edition*. Sas Inst, 1985.

[12] D. W. Hosmer Jr and S. Lemeshow, *Applied logistic regression*. John Wiley & Sons, 2004.

VITA

Hemanth Yadav Aradhyula was born in Guntur, India. He obtained his Bachelor of Technology (B.Tech) in Electrical and Communication Engineering from Koneru Lakshmaiah University, Guntur, in 2014. He did an internship with Haldex and two co-ops with Volvo. In May 2017, he obtained his Masters of Science (M.S) in Electrical Engineering from Missouri University of Science and Technology, Rolla. His focus areas included image processing, machine learning, and embedded systems.