



Computer Science and Engineering Discipline

Thesis Report

On

Smartphone-based Skin Cancer Detection using Image

Processing and Support Vector Machine.

Prepared By

Shagoto Rahman

ID: 170210

Computer Science and Engineering Discipline,

Khulna University.

Supervised By

Dr. Kamrul Hasan Talukder

Professor,

Computer Science and Engineering Discipline,

Khulna University.

Abstract

Cancer is one of the most lethal diseases around the globe. Among all the types of cancers, skin cancer is the most common and pernicious signifying the affected and death date. Our proposed thesis identifies such a problem and introduces a system of skin cancer detection. Image processing has been a turnaround for technology in recent years. The various processing stuff and feature mapping have gained efficiency in terms of computer vision. These features utilized by Support Vector Machine have brought a revolution in learning systems. Again, smartphones are cheaper these days. Keeping these in mind, a method has been proposed to detect skin cancer using image processing and support vector machine on the HAM10000 image dataset. We have compared our method with other skin cancer detection methods.

Contents

Introduction.....	7
1.1 Introduction.....	7
1.2 Problem Statement	8
1.3 Objectives	8
Background Topics	9
2.1 ABCD Rule.....	9
2.2 Support Vector Machine	10
Literature Survey	11
3.1 Introduction.....	11
3.2 Related Works.....	11
3.2.1 A Mobile Application for Early Detection of Melanoma by Image Processing Algorithms.....	11
3.2.2 Detection melanoma cancer using ABCD rule based on mobile device.....	14
3.2.3 Computer-Aided Early Detection and Classification of Malignant Melanoma.	16
3.2.4 m-Skin Doctor: A Mobile-Enabled System for Early Melanoma Skin Cancer Detection Using Support Vector Machine.....	17
Proposed Method	20
4.1 Introduction.....	20
4.2 Methodology	20
4.2.1 Pre-processing.....	22
4.2.2 Segmentation.....	23
4.2.3 Feature Extraction	24
4.2.4 Feature Selection and Classification	28
4.2.5 Training.....	29
4.2.5 Testing.....	29
Dataset	30
Result and Discussion	31
6.1 Introduction.....	31
6.2 Accuracy Matrices	31
6.2.1 Classification Accuracy	31
6.2.2 Confusion Matrix	31
6.2.3 Precision-Recall-Specificity-F1-Score.....	32
6.2.4 ROC Curve.....	32

6.3 Result	32
6.3.1 Feature Selection.....	32
6.3.2 Accuracy	38
6.3.3 Precision-Recall-Specificity-F1-Score.....	39
6.3.4 ROC Curve.....	40
6.4 Comparison.....	42
Implementation	43
Conclusion	46
8.1 Conclusion	46
8.2 Future Work Direction.....	46
Bibliography	47

List of Figures

Fig. 2. 1: Upper and lower parts of an image over the x-axis.	9
Fig. 2. 2: Support Vector Machine mechanism.	10
Fig. 3. 1: Overall workflow of [8].	12
Fig. 3. 2: Preprocessing with Gaussian and Otsu.	12
Fig. 3. 3: Extra use of Laplacian filter.	13
Fig. 3. 4: Overall methodology of the authors [6].	14
Fig. 3. 5: Overall work [9].	16
Fig. 3. 6: Segmentation process.	17
Fig. 3. 7: Overall Work [12].	18
Fig. 3. 8: (a) Lesion selection by the user (b) m-Skin is applying the grab cut algorithm (c) segmented image after applying Grab Cut algorithm.	18
Fig. 4. 1: Block diagram.	20
Fig. 4. 2: Flowchart of the proposed method.	21
Fig. 4. 3: Pre-processing.	23
Fig. 4. 4: Segmentation.	24
Fig. 4. 5: Biggest skin blob finding and ellipse fitting.	25
Fig. 4. 6: Asymmetry.	26
Fig. 4. 7 : SVM structure.	28
Fig. 6. 1: Dataset after feature extraction.	33
Fig. 6. 2: Feature ranking using Chi-Square.	34
Fig. 6. 3: Accuracies of features chunks using SVM on 80% training and 20% testing.	34
Fig. 6. 4: Accuracies of features chunks using SVM of 90% training and 10% testing.	35
Fig. 6. 5: Correlation of features with target data.	36
Fig. 6. 6: Feature ranking after discounting 'irc' and 'ird'.	37
Fig. 6. 7: Accuracy of the features chunks in the correlation analysis using SVM on 80% training, 20% testing.	37
Fig. 6. 8: Accuracy of the features chunks in the correlation analysis using SVM on 90% training, 10% testing.	38
Fig. 6. 9: Performance measures of SVM on 80% training, 20% testing.	39
Fig. 6. 10: Performance measures of SVM on 90% training, 10% testing.	40
Fig. 6. 11: ROC Curve of SVM on 80% training, 20% testing data.	41
Fig. 6. 12: ROC Curve of SVM on 90% training, 10% testing data.	41
Fig. 7. 1: Detection of malignant melanoma in two different smartphones.	44
Fig. 7. 2: Detection of benign melanoma in two different smartphones.	45

Tables

Table 3. 1: ACCURACY, SPECIFICATION, AND SENSITIVITY OF PROPOSED APPLICATION...	13
Table 3. 2: Score and Weight factor of the features.....	15
Table 3. 3: Prediction by TDS	15
Table 3. 4: Different results on two phones.	16
Table 3. 5: Accuracy results achieved by m-Skin Doctor.....	19
Table 4. 1: RGB ranges for the six colors [18].	27
Table 6. 1: Confusion matrix.	32
Table 6. 2: Confusion matrix (80% Testing training, 20% testing)	38
Table 6. 3: Confusion matrix (90% Testing training, 10% testing)	39
Table 6. 4: Precision-Recall-Specificity-F1-Score of the test data.	39
Table 6. 5: Comparison with other methods.	42

Chapter 1

Introduction

1.1 Introduction

Cancer is treated as one of the life-threatening diseases found in humans. It is the uncontrolled growth of abnormal cells anywhere in a body. Cancers are of many types depending on the areas they converge and lung, breast, colon, prostate, ovarian, skin are the common ones. Among all of them, the most striking one is skin cancer [1]. Malignant melanoma is defined as the most lethal type of skin cancer. More than 9,500 people are being diagnosed with skin cancer daily in the United States where the death rate is two people per hour [2]. This gives alarming statistics about the havoc skin cancer can make. One good news is that, when detected early, the 5-year survival rate increases to 99 percent [2]. So, the mortality rates among skin cancer-affected patients can be decreased if it is diagnosed early. Existing technologies express the need for cell removal techniques from the affected area that is technically known as ‘biopsy’ [3]. This biopsy is not only tedious for dermatologists but also painful and costly for the patients as well. Moreover, the time consumption in the biopsy, as well as report generation, is huge. Medical professionals are mostly needed for this purpose and as a result people in remote places, as well as places where the executives are in short in number, indicate serious deprivation. That is why a system that will classify whether a cell is cancerous or not in seconds efficiently is mostly needed. Furtherance of technology and devices has enabled the diagnosis of diseases efficient which is popularly known as Computer-Aided Diagnosis systems [4]. Skin cancer detection has been alleviated strongly by these systems recently and they have nearly replaced the biopsy systems in terms of cost and time consumption. Various techniques have been engendered by these computer-aided systems. One of the most practiced and efficient is the ABCD [5] rule. Image processing has an unmatched role here. The image of the affected area is used to detect various features that further help for classification. The recent advancement in image processing has brought revolutionary changes in feature map generation. Next, these feature maps are learned with various learning algorithms to classify skin cancer. But still, the problem remains in terms of cost as computers are not feasible for people of all walks. To reach a large amount of people computers cannot be the only savior. So as the cameras since taking a picture and then processing in the computer take further time. Smartphones in these situations can come as a rescue since cameras are embedded in smartphones. Moreover, smartphones are not costly these days, and almost in every house, there is a smartphone. So, a quick solution can be this if a skin cancer detection scheme is embedded with smartphones. The challenge would be applying image processing algorithms in smartphone devices and building a system that detects a skin lesion to be cancerous or not.

1.2 Problem Statement

The alarming affected and the death rate provides information about the lethal features of skin cancer. Early detection can palliate the situation and give a better life to humanity. Again, among all unfeasible platforms and devices to people of all walks, skin cancer detection with a smartphone using images can be very handy in this situation.

The main problem is to detect melanoma skin cancer in smartphones using images.

1.3 Objectives

The main object of our work is to detect skin cancer in smartphones using images of the affected area.

Chapter 2

Background Topics

2.1 ABCD Rule

ABCD rule refers to using the ABCD features in dermatology namely A for Asymmetry. B for Border, C for Color, and D for diameter [6]. Asymmetry refers to the nonuniformity of skin lesions over two axes. The nonoverlapping part over x-axis and the nonoverlapping part over the y-axis are averaged to measure the entire asymmetry of the skin lesion. Fig 2.1 shows the two parts of an image over the x-axis. Both of the partial images are overlapped to get the uncommon part over the x-axis and the same goes for the y-axis to get the average asymmetry value.

Border refers to the dermoscopic feature that reflects at the irregular border of skin lesions. The border is referred to as the circularity index [6]. The circularity index is calculated as the following equation:

$$CI = 4A\pi / P^2 \quad (2.1)$$

Where, CI= Circularity Index, A= Area of Skin Lesion, P =Perimeter.

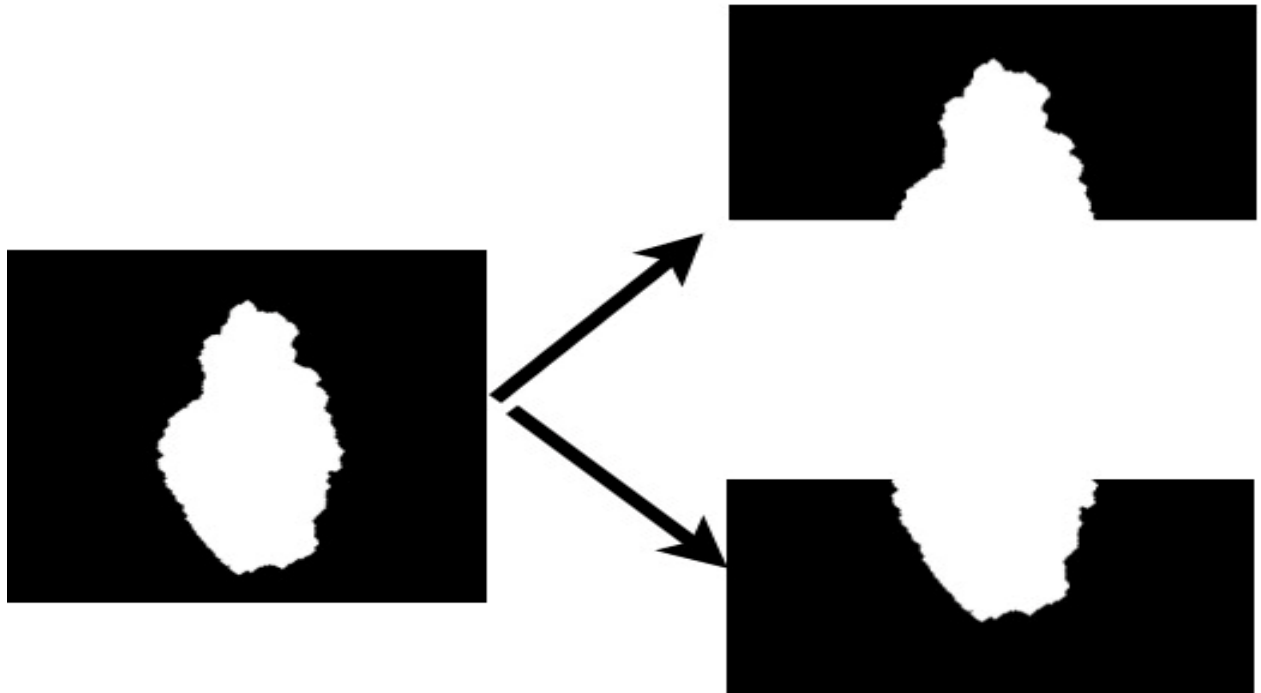


Fig. 2. 1: Upper and lower parts of an image over the x-axis.

Another important feature is Color. A malignant skin is supposed to have 4 to 6 of the 6 main colors namely white, black, red, light brown, dark brown, blue-gray. However benign cell is supposed to have none to 3 of these colors.

The last feature in this chunk is the Diameter. If the Diameter of a skin lesion is more than 6 mm then it has a significant consideration of being malignant melanoma.

All of these features are required to calculate the total dermoscopy score TDS and that is:

$$\text{TDS} = A * 1.3 + B * 0.1 + C * 0.5 + D * 0.5 \quad (2.2)$$

If this TDS score is greater than 5.75 then there are higher chances of being the lesion as a malignant one.

2.2 Support Vector Machine

Support Vector Machine (SVM) [7] is one of the most widely used machine learning algorithms all over the globe. SVM is such kind of a machine learning algorithm that works by creating an optimal hyperplane that separates the target classes. This significant operation is done using the support vectors, vectors that are used to represent the optimal hyperplane. Support vectors are chosen from each of the attributes and they are often regarded as the closest to the optimal hyperplane and the hyperplane is created in such a mechanism as it is at the maximum distance from all the support vectors. Fig 2.2 shows the support vector machine mechanism.

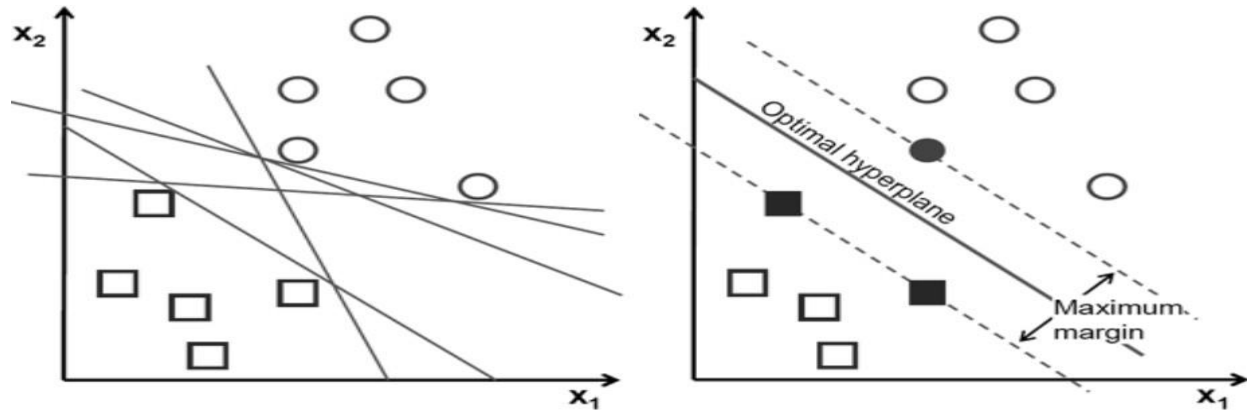


Fig. 2. 2: Support Vector Machine mechanism.

Chapter 3

Literature Survey

3.1 Introduction

We have engendered a survey on skin cancer detection in smartphone devices using images. Following websites have been queried for the purpose:

- ieeexplore.ieee.org (IEEE Xplore Digital Library)
- www.sciencedirect.com
- link.springer.com
- dl.acm.org (ACM Digital Library)
- scholar.google.com

A lot of papers have been found on skin cancer detection using images but in terms of smartphones very few papers are there. We started the searching for skin cancer, then went into the mainstream by specifically pointing at smartphone-based systems. Getting the results from the query we selected the relevant papers by extracting the meaning of the title and abstract of the papers. After that regarding our literature survey, we studied various papers.

3.2 Related Works

3.2.1 A Mobile Application for Early Detection of Melanoma by Image Processing Algorithms.

Alizadeh et al. proposed a method to detect melanoma using image processing techniques in a mobile application [8]. The authors built a smartphone application to detect malignant melanoma. Their methodology was based on image processing to find out important features of a skin lesion to determine cancerous or not. The methodology commences with the acquisition of images using smartphones. Fig 3.1 depicts the methodology of the authors.

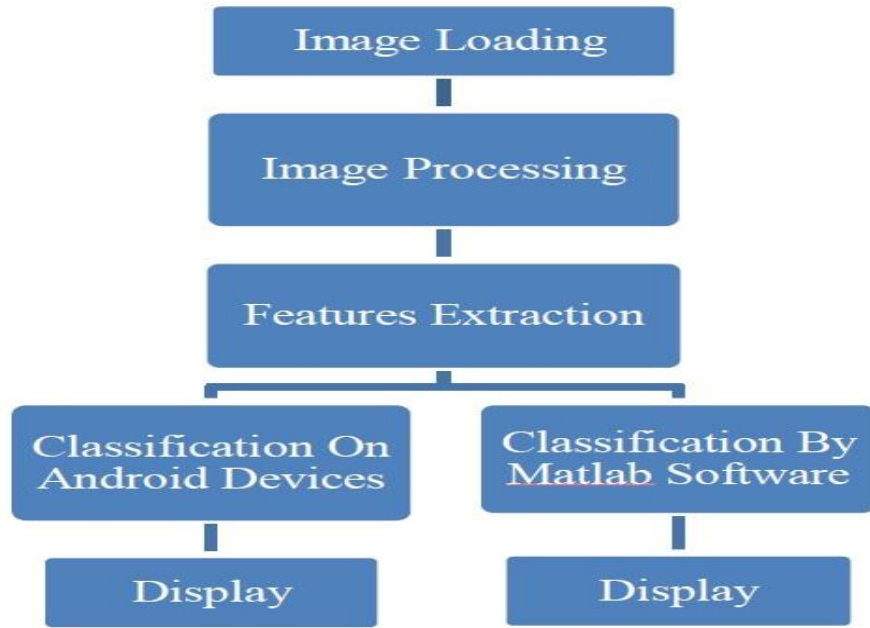


Fig. 3. 1: Overall workflow of [8].

The preprocessing of images was done using a Gaussian filter to reduce noise. Many of the images did not need any noise-canceling operation as they were noise-free. For segmentation of the images, Otsu's method had been used. In addition, a Laplacian filter was used for those images that were not segmented fully using Otsu's thresholding method. Figures 3.2 and 3.3 illustrate the use of Gaussian and Laplacian filters with Otsu's thresholding respectively.

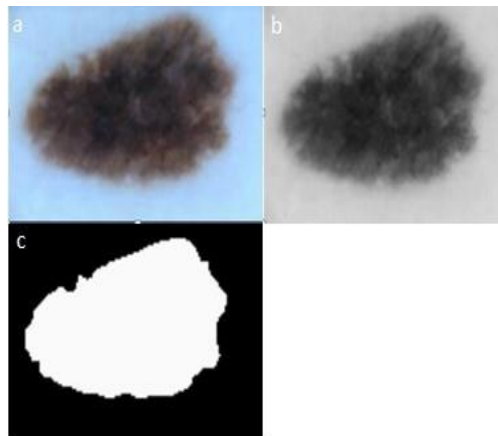


Fig. 3. 2: Preprocessing with Gaussian and Otsu.

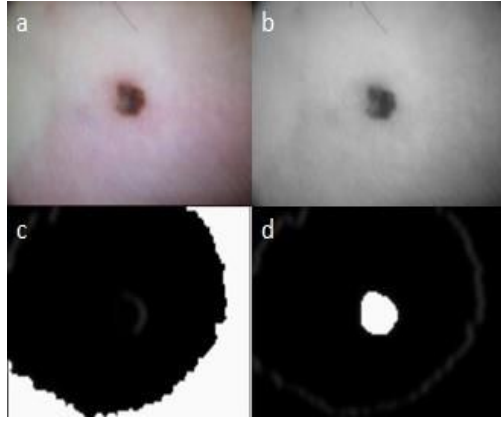


Fig. 3. 3: Extra use of Laplacian filter.

As for features, the authors used the following features:

- **Irregularity Index A** = P/A
- **Irregularity Index B** = P/MJ
- **Irregularity Index C** = $P(\frac{1}{MI} - \frac{1}{MJ})$
- **Irregularity Index D** = $MJ-MI$
- **Circulatory Index, CI** = $4A\pi / P^2$

Where A= Area, P= Perimeter, MJ= Major Index, MI= Minor Index.

Next, the authors went further with the normal Bayesian method for smartphones and SVM for the computer-based system to compare the performance between them on 150 personally collected images. The smartphone-based approach achieved 93% overall accuracy and the computer-based approach achieved 96% accuracy. Table 3.1 shows the performance comparison of these two approaches. The method works well with only geometrically distinct images.

Table 3. 1: ACCURACY, SPECIFICATION, AND SENSITIVITY OF PROPOSED APPLICATION.

Methods	SVM (MATLAB)	Normal Bayesian (Smart phone)
Accuracy of Melanoma Detection	93.33%	90%
Accuracy of Non-melanoma Detection	100%	96.67%
Overall Accuracy	96.67%	93.33%
Overall Sensitivity	100%	96.43%
Overall Specificity	93.75%	90.63%

3.2.2 Detection melanoma cancer using ABCD rule based on mobile device.

Firmansyah et al. proposed a method to detect melanoma using the ABCD rule on the mobile device [6].

The authors stated about an android application to detect malignant melanoma using ABCD (Asymmetry, Border, Color, Diameter) rule. Fig 3.4 depicts the overall workflow of the paper.

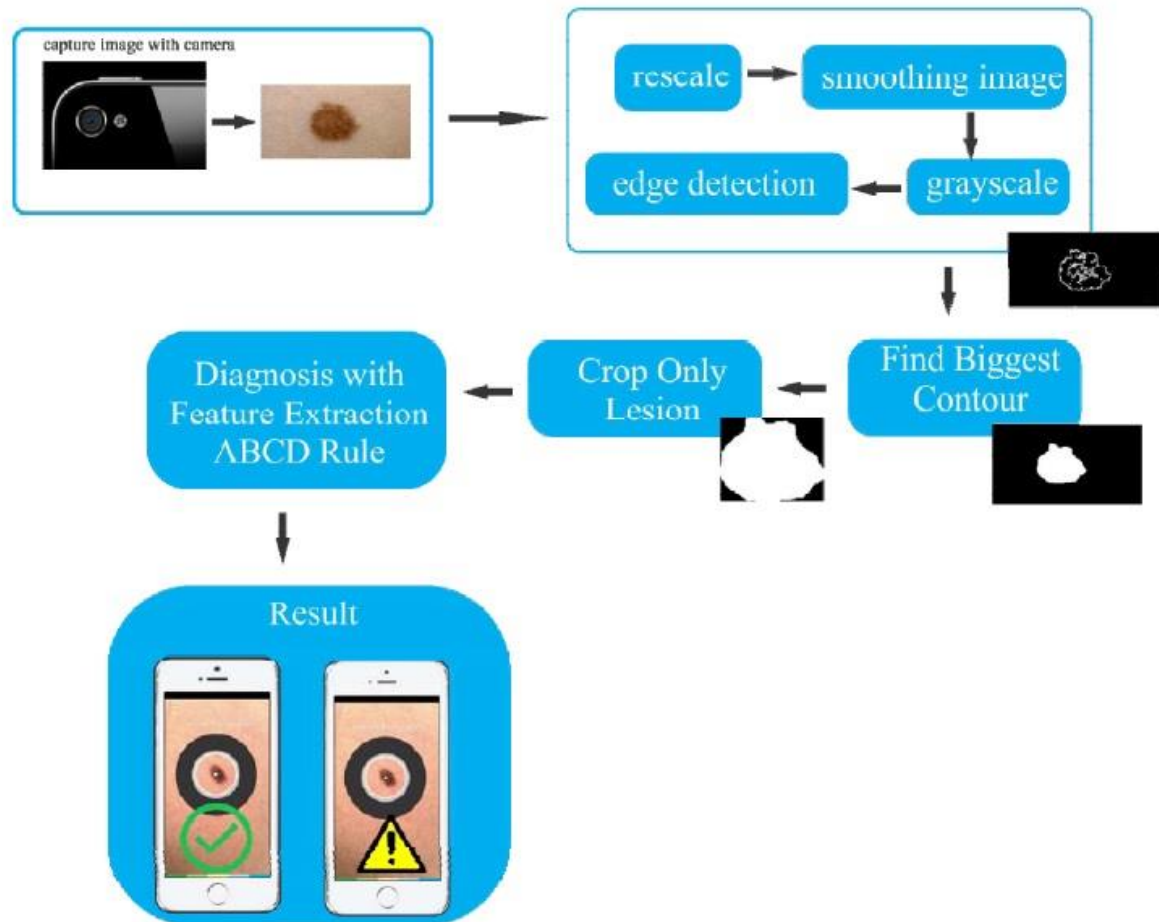


Fig. 3. 4: Overall methodology of the authors [6].

After image acquisition, the method proceeded with resizing the image and smoothing using the median filter. They used the canny edge detection procedure to detect the edge of skin lesions and segmentation. As for features they have used ABCD (Asymmetry, Border, Color, Diameter). After the TDS (Total Dermoscopy Score) generation, they decided whether it is melanoma or not.

For the dataset, the authors had used a digital clinical image medical dataset from PH2 - A dermoscopic image database for research and benchmarking and the International Society for Digital Imaging of the Skin with histologic diagnosis. There were 40 images for testing specifically 20 images benign and 20 melanoma.

The score and weight factor of each feature are given in Table 3. 2:

Table 3. 2: Score and Weight factor of the features

Criteria	Score	Weight Factor
Asymmetry(A)	0-2	1.3
Border(B)	0-8	0.1
Color(C)	1-6	0.5
Diameter(D)	1-5	0.5

Based on the values the Total Dermoscopy Score (TDS) was calculated as:

$$TDS=1.3*A+0.1*B+0.5*C+0.5*D$$



The prediction was based on the following Table 3.3:

Table 3. 3: Prediction by TDS

Total Dermoscopy Score (TDS)	Classification
<4.75	Benign Melanoma
4.75-5.75	Suspicious Skin
>5.75	Malignant Melanoma

The method showed different results for different smartphones. Table 3.4 shows the same phenomenon showcasing different results in two different smartphones Redmi and Samsung.

Table 3. 4: Different results on two phones.

Different Crop Result	
	
Redmi Note 2	Samsung Tab S
TDS= 4.93	TDS=4.73

3.2.3 Computer-Aided Early Detection and Classification of Malignant Melanoma.

Shafiq et al. proposed a computer-aided method for the same purpose [9]. Fig 3.5 shows the current best solution of the paper. Image pre-processing was ensured by the dull razor algorithm. Dull razor algorithm eradicated all human hair noise and other artifacts.

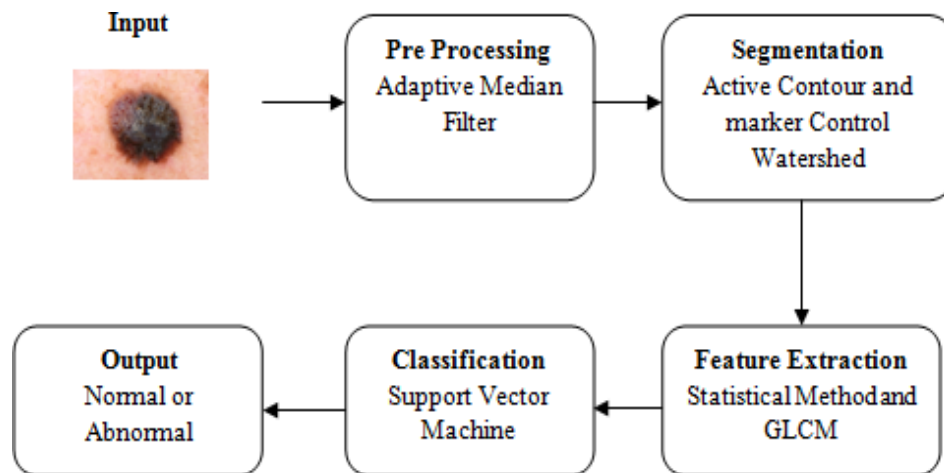


Fig. 3. 5: Overall work [9].

For segmentation, an automatic segmentation process was used. The ROI (Region Of Interest) is determined by it. Then the boundary of ROI was identified and they had separated the background and foreground. Figure 3.6 shows the segmentation of the method.

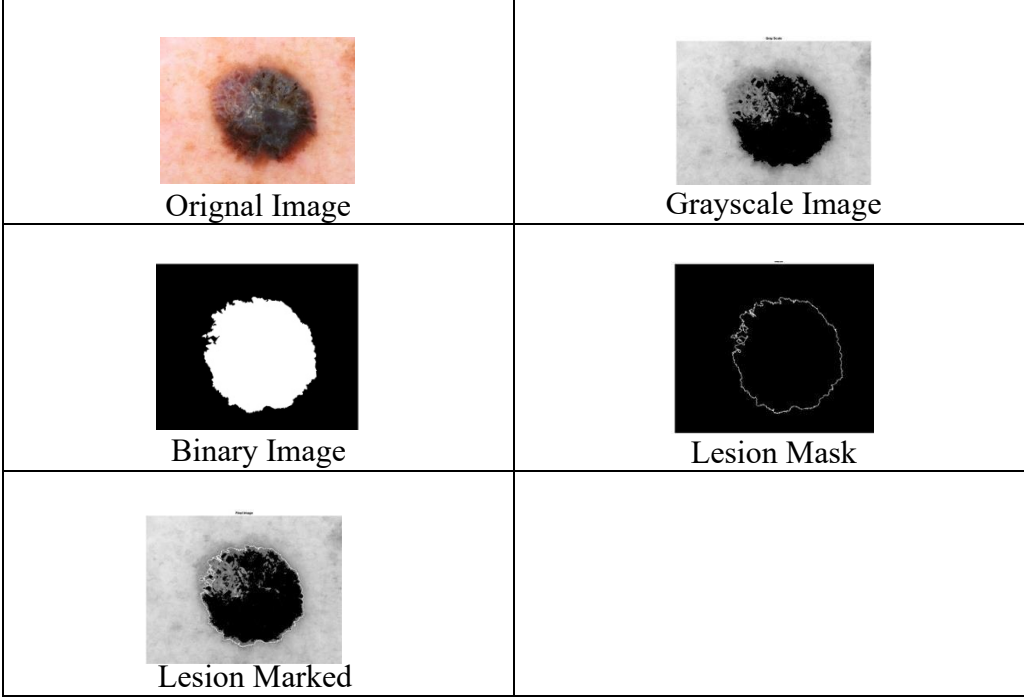


Fig. 3. 6: Segmentation process.

As for features they had used both GLCM (Grey Level Co-occurrence Matrix) and statistical features. As for statistical features area, mean, variance, standard deviation had been used. From GLCM features that were used are contrast, homogeneity, local homogeneity, dissimilarity, entropy, correlation, shade, sum average, sum entropy, cluster prominence, and difference entropy. The dataset of the authors consisted of 50 benign and 50 melanoma images attained from DermNet [10] and the ISIC archive [11].

The classification was done using SVM (Support Vector Machine). The classification accuracy was defined as 89% where sensitivity and specificity were 86.9% and 91.8% respectively.

3.2.4 m-Skin Doctor: A Mobile-Enabled System for Early Melanoma Skin Cancer Detection Using Support Vector Machine.

Another mobile-based skin cancer detector named “m-Skin Doctor” was addressed by authors in their paper [12]. Fig. 3.7 shows the overall work process of the system.

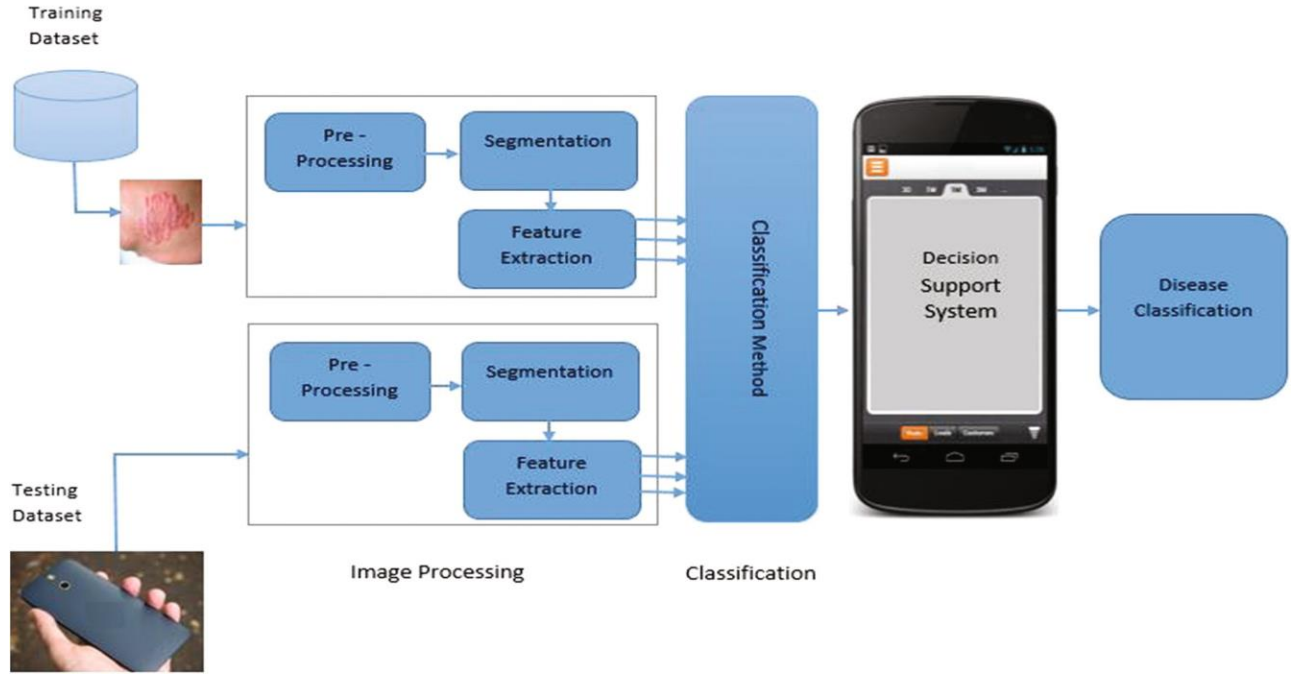


Fig. 3. 7: Overall Work [12].

The preprocessing of the images was engendered by giving the images a dimension of 640x480 along with the Gaussian filter to remove noise. Grab Cut algorithm had been enabled to segment images into four sections namely Exactly Background, Probably Background, Exactly Foreground, and Probably Foreground. Figure 3.8 shows the segmentation process where the segmentation area was selected by the user.

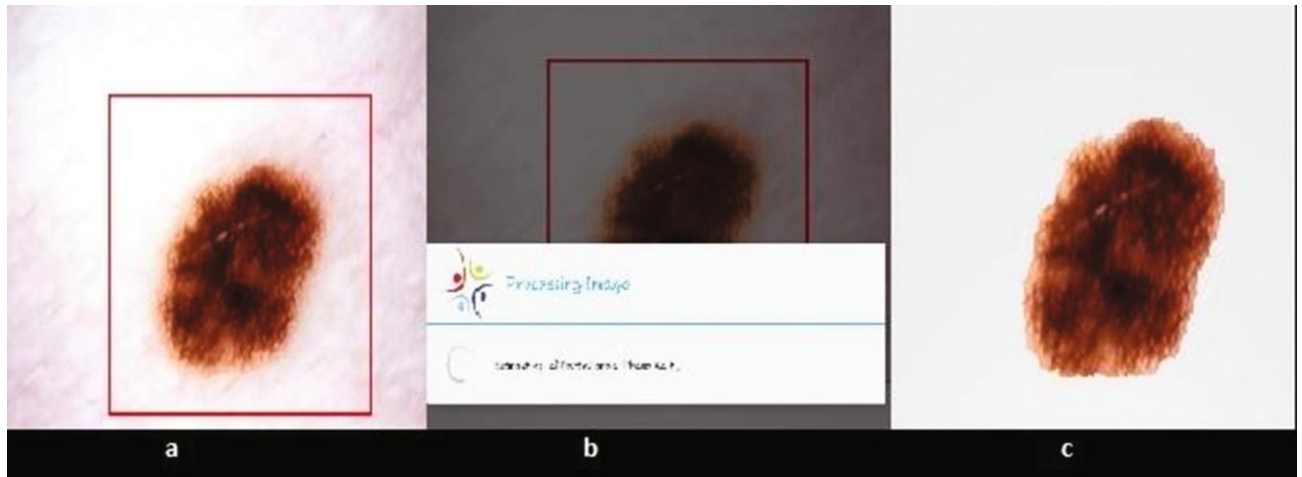


Fig. 3. 8: (a) Lesion selection by the user (b) m-Skin is applying the grab cut algorithm (c) segmented image after applying Grab Cut algorithm.

After segmentation the method propagated with features extraction. The basic features enabled in the section were area, perimeter, eccentricity, mean, standard deviation, L1 norm, L2 norm angle of lesion, major and minor axis of the lesion. Using SVM as the classifier the method achieved

sensitivity and specificity of 80% and 75% respectively. Table 3.5 shows the results table.

Table 3. 5: Accuracy results achieved by m-Skin Doctor.

	Melanoma	Nonmelanoma
Melanoma	80%	20%
Nonmelanoma	25%	75%

Chapter 4

Proposed Method

4.1 Introduction

There are various methods in skin cancer detection. We have enabled a novel system of classifying skin cancer in smartphone. In this chapter, we will showcase our proposed method, steps of the experiment using block diagram, training, and testing.

4.2 Methodology

In our proposed system, we have advocated a method to detect melanoma skin cancer in smartphones using skin lesion images. For that purpose, we have enabled various features both geometrical and color with the Support Vector Machine (SVM) classifier. The dataset that has been used in this regard is the HAM10000 dataset [13]. Fig 4.1 and 4.2 illustrate the block diagram and the flowchart of our proposed system respectively. Our proposed study is cleaved into five chunks. They are:

1. Pre-processing
2. Segmentation
3. Feature Extraction
4. Feature Selection and Classification
5. Assigning classification parameters in Smartphone.

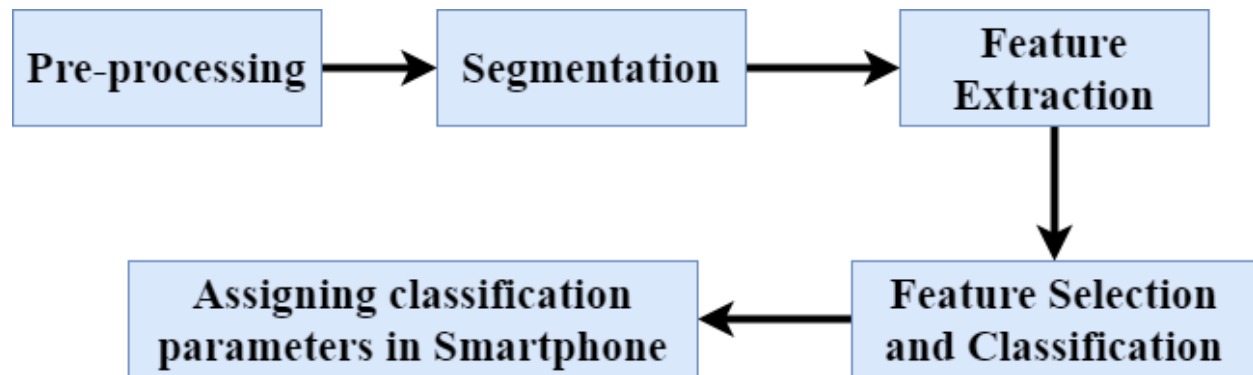


Fig. 4. 1: Block diagram.

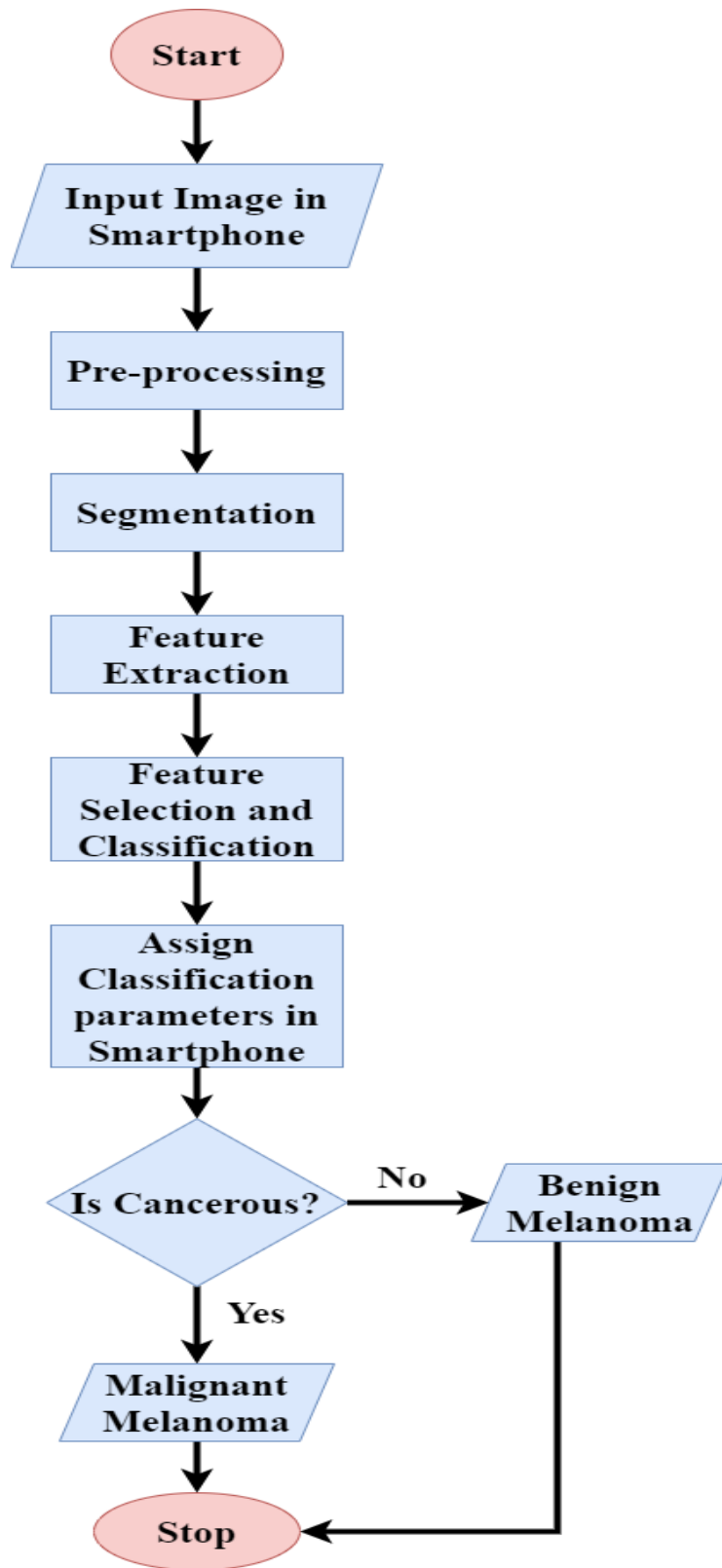


Fig. 4. 2: Flowchart of the proposed method.

4.2.1 Pre-processing

We have two different types of features in this method, one is the geometrical features and the other is the feature related to colors. For that purpose, we have initiated the method by preprocessing the images. First, we have resized the images and removed noise from the images. Next, we have eradicated the human hair noises from the images as well. Thus the pre-processing is divided into three steps (1) Resizing, (2) Noise Removal, (3) Hair Removal. All of them are done in smartphone.

Resizing: The images have been resized into 600 x 600-pixel size. Different images have different resolutions. To make them into the same size this resizing procedure has been advocated.

Noise Removal: For preprocessing of the images, we have used the Gaussian filter [14]. The Gaussian filter is that kind of filter that uses the Gaussian function. The function is defined as follows:

$$g(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (4.1)$$

The function creates such kind of a filter or kernel where there are higher weights in the middle and low values at the edges. Using that kernel the image will be blurred. It ensures higher intensity at the middle and again the edges will be clear as well. The Gaussian filter removes higher frequencies from the images and passes the low frequencies.

Hair Removal: We have utilized closing to remove hair noises. Closing is a morphological operation. It is defined simply as a dilation followed by erosion using the same structuring element for both operations [15]. Dilation adds a layer of pixels to both the inner and outer boundaries of regions. Erosion slims the image by removing a layer of pixels from the boundaries. The holes and gaps between different regions become larger, and small details are eliminated. So, closing is denoted by,

$$A \bullet B = (A \oplus B) \ominus B \quad (4.2)$$

Where, A is the image and B is the kernel that will be applied to the image to do the operations.

That is why we would use the closing operation so that the image would first be joined with all sorts of its connected parts using dilation and then all the parts that are shallowly connected would get cleaned by erosion which will ensure the eradication of human hair and other artifacts from images.

Fig. 4.3 depicts the preprocessing of an image.

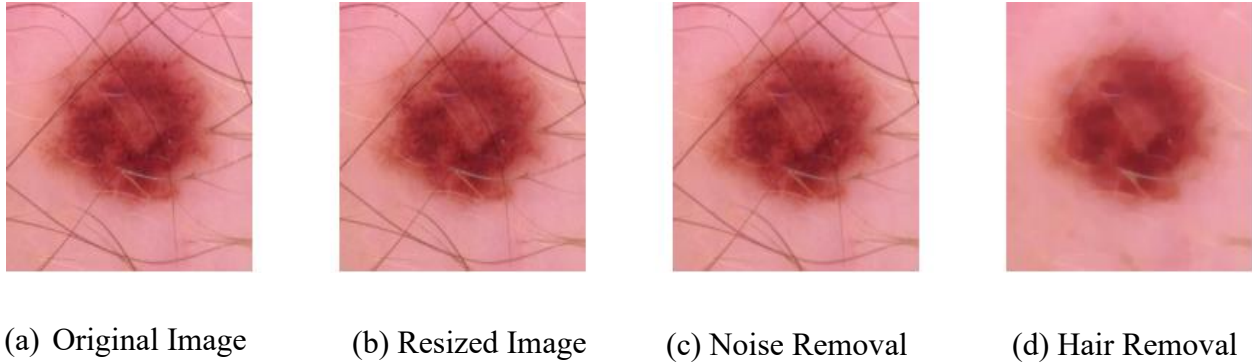
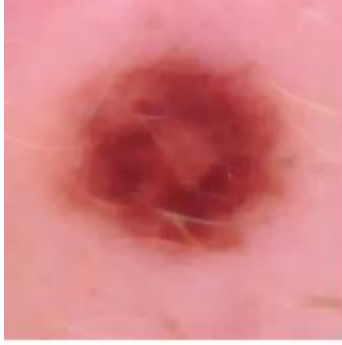


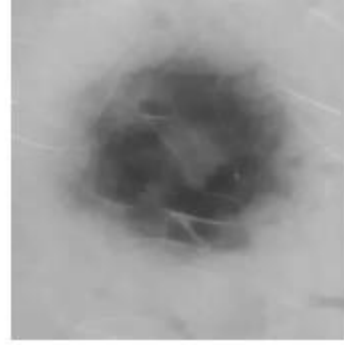
Fig. 4. 3: Pre-processing.

4.2.2 Segmentation

From our research, we have learned that Otsu's thresholding would be the best method for segmentation in this context [16]. Otsu's thresholding iterates through all the greyscale levels present in the image to find the desired threshold and calculate the intraclass variance and inter-class variance to select the best value to divide foreground and background. Thus Otsu's thresholding discriminates the lesion area from the background. Otsu's thresholding works fine with enhanced image edges. This algorithm will compute intraclass and inter-class variances and find out the best threshold that has the least intraclass and highest inter-class variance, which would purely segment the lesion area from its background. This procedure is also done on a smartphone. Fig 4.4 illustrates the segmentation proceedings.



(a) Preprocessed Image



(b) Gray Image



(c) Otsu's thresholding



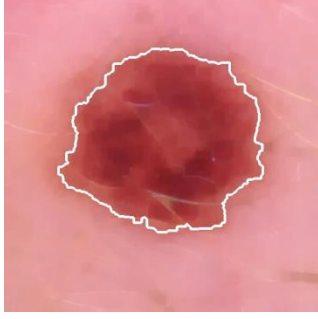
(d) Bitwise not

Fig. 4. 4: Segmentation.

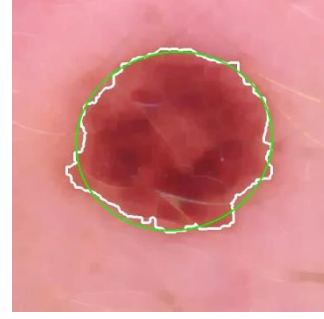
4.2.3 Feature Extraction

4.2.3.1 Geometrical features

The geometrical features are related to the area, perimeter, and major and minor axis of the skin area [17]. For getting the major and minor axis, we have fitted an ellipse within the segmented area. Fig 4.5 portrays the ellipse fitting phenomenon.



(a) Largest skin blob



(b) Fitting ellipse

Fig. 4. 5: Biggest skin blob finding and ellipse fitting.

For the geometrical features, we would need some terms to be introduced. They are:

A= Area of segmented lesion, P= Perimeter of segmented lesion.

MJ= Major Axis. MI=Minor axis.

So, these features are:

- **Irregularity Index A** = P/A (4.3)

- **Irregularity Index B** = P/MJ (4.4)

- **Irregularity Index C** = $P \left(\frac{1}{MI} - \frac{1}{MJ} \right)$ (4.5)

- **Irregularity Index D** = $MJ-MI$ (4.6)

- **Circulatory Index, CI** = $4A\pi / P^2$ (4.7)

- **Asymmetry Index** = $(\Delta Ax + \Delta Ay) / (2 * A)$ (4.8)

Where, ΔAx = Pixel difference over x axis,

ΔAy = Pixel difference over y axis.

Fig. 4.6 illustrates the asymmetry.

- **Edge Abruptness** =
$$\frac{\frac{1}{P_L} \sum_{p \in C} (d_R(p, G_L) - m_d)^2}{m_d^2}$$
 (4.9)

Where, P_L = Perimeter.

G_L = Center of the skin blob.

m_d = mean distance of the boundary points.

C= Boundary points.
 d_R = Distance between two points.



(a) Thresholded Image



(b) Straightened Image



(c) Upper part over X-axis



(d) lower part over X-axis



(e) XOR over X-axis



(f) 90 degree rotation of (b)



(g) Upper part of (f) over new X axis.



(h) Lower part of (f) over new X axis.



(i) XOR of (f) over new X axis.

Fig. 4. 6: Asymmetry.

4.2.3.2 Color features

4.2.3.2.1 Colors

Another important feature regarding skin cancer detection is the number of colors that are pigmented in a skin lesion. Six colors are regarded in this particular feature detection. They are white, black, red, light brown, dark brown, blue-gray. Again, malignant melanoma is supposed to have 4 to 6 of these colors on the other hand a benign cell is said to have from 0 to 3 of these colors usually. A color is said to be present in a skin lesion if the number of the particular color in the skin lesion exceeds 0.1% of the total pixels. Table 4.1 shows the RGB ranges [18] of the six colors that have been used in this proposed method.

Table 4. 1: RGB ranges for the six colors [18].

Color	RGB ranges
Black	$(R \leq 62), (G \leq 52), (B \leq 52)$
White	$(R \geq 205), (G \geq 205), (B \geq 205)$
Red	$(R \geq 150), (G < 52), (B < 52)$
Light-Brown	$(150 \leq R \leq 240), (50 < G \leq 150), (0 \leq B \leq 100)$
Dark-Brown	$(62 < R < 150), (0 \leq G < 100), (0 < B < 100)$
Blue-Gray	$(0 \leq R \leq 150), (100 \leq G \leq 125), (125 \leq B \leq 150)$

4.2.3.2.2 Mean

Mean of all the pixels in the lesion. It is derived by the following equation:

$$\text{Mean} = \sum_{i=0}^{\text{height}-1} \sum_{j=0}^{\text{width}-1} f(i, j) / N \quad (4.10)$$

4.2.3.2.3 Standard Deviation

Standard deviation is another color feature that is defined by:

$$\text{Std} = \sqrt{\sum_{i=0}^{\text{height}-1} \sum_{j=0}^{\text{width}-1} (f(i, j) - \text{Mean})^2 / N} \quad (4.11)$$

4.2.4 Feature Selection and Classification

4.2.4.1 Support Vector Machine

All of the features are extracted using a smartphone. Now all of these 10 features would be used for classification in computer. According to our research, SVM (Support Vector Machine) [7] is the best approach for classification. Classifiers like Bayesian classifiers [19], decision tree [20] are not as efficient as SVM in this context. SVM performs better with both continuous features and large datasets. A lot of betterments have been done in SVM in recent times [21].

SVM is a machine learning approach. SVM constructs hyper planes defining decision boundary of classification. SVM creates optimal hyperplane by defining support vectors and using those support vectors SVM obtains the optimal that has the best distance among the support vectors. Following equations show the same.

$$W^T * X - b = +1 \quad (4.12)$$

$$W^T * X - b = -1 \quad (4.13)$$

Where, X = feature vector, W = weight vector, b = Bias, $+1$ = Positive Class, -1 = Negative Class.

$$\text{Optimal Hyperplane} = W^T * X - b = 0. \quad (4.14)$$

Figure 4.7 shows the SVM structure.

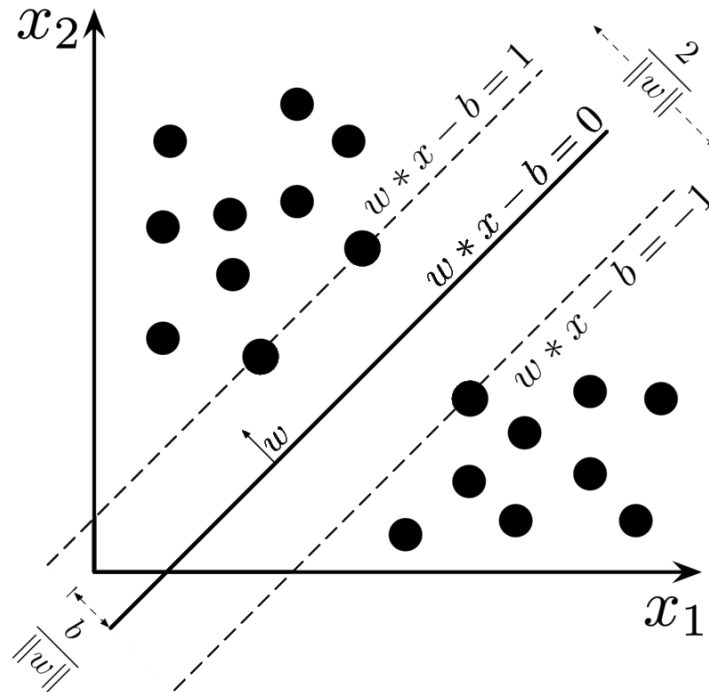


Fig. 4. 7 : SVM structure.

4.2.4.2 Feature Selection

Chi-square and Correlation have been enabled in this section to witness any relation among the variables as well as target class.

4.2.4.2.1 Chi-square

A Chi-square test is done to check the dependency between variables. The null hypothesis for the chi-square test is considered as there is no relationship between variables, whereas the alternative hypothesis is considered that there is a significant relationship among the variables. Chi-square value between two variables can be extracted as the following equation:

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (4.15)$$

Where, X^2 = Chi-Square value.

O= Observed value.

E= Expected value if there was no relationship between the variables.

The higher chi-square value denotes the dependency between variables.

4.2.4.2.2 Correlation

Another method to know the relationship among the variables is the correlation. The correlation between two variables signifies the linear relationship among them. Correlation between two variables can be inferred from the following equation:

$$\text{Correlation} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (4.16)$$

Where, x= Feature one, y= Feature two.

The correlation value range is $[-1, +1]$, where -1 denotes a highly negative relationship and +1 denotes a highly positive relationship and 0 denotes no relationship. After feature selection and classification the learning parameters of the SVM will be merged with the smartphone for prediction.

4.2.5 Training

For training of the dataset, we have engendered 80% and 90% training instances respectively.

4.2.5 Testing

For testing of the dataset, we have engendered 20% and 10% testing instances respectively.

Chapter 5

Dataset

HAM10000 [13] dataset has been enabled for the research perspective. The dataset contains 10015 dermoscopic images of which 1113 malignant lesion images and 1099 are benign lesions. Using histopathology 50% of lesions are confirmed as well as ground truth for the other cases is confirmed by follow-up, expert consensus, or confocal microscopy.

Chapter 6

Result and Discussion

6.1 Introduction

The proposed work is to recognize skin cancer from images. In the HAM10000 dataset we have two classes, one is malignant (cancerous) and the other is benign (non-cancerous).

6.2 Accuracy Matrices

There are various matrices in terms of model evaluation. We have used some of them to evaluate our model as well.

6.2.1 Classification Accuracy

Classification accuracy is defined by the ratio of the number of correct predictions to the total number of input samples. So, the accuracy of a model is defined by the following equation:

$$\text{Accuracy} = \frac{\text{Number of Correctly Classified Samples}}{\text{Total number of samples}} \quad (6.1)$$

6.2.2 Confusion Matrix

Another important model evaluation process is determining the confusion matrix. A summary of predictions of a particular model is presented by a confusion matrix. Moreover, the entries of the matrix represent different prediction statistical results of the model. The size of a confusion model is defined as $p \times p$ where p is the number of target classes and p is at least greater than or equal to 2. Now four important terms indulge in the generation of the confusion matrix. They are:

True Positives (TP): The cases where the prediction is positive and the prediction is correct.

True Negatives (TN): The cases where the prediction is negative and the prediction is correct.

False Positives (FP): The cases where the prediction is positive and the prediction is incorrect.

False Negatives (FN): The cases where the prediction is negatives and the prediction is incorrect.

Now we can define the confusion matrix as Table 6.1.

Table 6. 1: Confusion matrix.

	Positive	Negative
Positive	TP	FN
Negative	FP	TN

6.2.3 Precision-Recall-Specificity-F1-Score

Precision: Precision refers to the proportion of actual positive tuples among all positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6.2)$$

Recall: Recall corresponds to the proportion of correctly classified positive tuples among all positive tuples.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6.3)$$

Specificity: Specificity refers to the ratio of correct negative predictions among the negative tuples.

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (6.4)$$

F1-Score: F1-Score is the harmonic mean of precision and recall.

$$\text{F1-Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (6.5)$$

6.2.4 ROC Curve

ROC (Receiver Operating Characteristic) curve is the graphical representation of the performance of a model by plotting the true positive rates versus the false positive rates of the prediction of a model. AUC (Area Under Curve) refers to the 2D area underneath the ROC curve that represents the performance of the model.

6.3 Result

6.3.1 Feature Selection

After getting all the features, we have enabled the chi-square and correlation to establish a set of features that will approximately triumph the best results. Fig. 6.1 illustrates the dataset after feature extraction, where,

- **ira** = Irregularity Index A,
- **irb** = Irregularity Index B,

- **irc** =Irregularity Index C,
- **ird** =Irregularity Index D,
- **ci** =Circularity Index,
- **ai** = Asymmetry Index,
- **edge** = Edge Abruptness,
- **col** = Number of colors,
- **mean** = Mean of the colors,
- **std** = Standard Deviation of colors.
- **label** = (0: benign, 1: malignant).
- **name** = name of image.

The chi-square statistic between the variables and the target feature has been determined to indicate the target feature dependency with the features. For that purpose, the features were ranked using the chi-square values with the target feature 'label'. Fig. 6.2 depicts the feature ranking using chi-square.

	ira	irb	irc	ird	ci	ai	col	edge	mean	std	label	name
0	0.017561	3.535974	2.491174	190.741180	0.438542	0.122329	3	0.037511	97.774860	23.536211	0	ISIC_0024306.jpg
1	0.025185	4.987564	2.796229	160.481630	0.223943	0.243822	1	0.074869	127.018330	13.976082	0	ISIC_0024307.jpg
2	0.013661	3.090491	3.414616	329.705720	0.473869	0.086903	3	0.066670	92.389990	22.156944	0	ISIC_0024308.jpg
3	0.016668	3.828649	1.237170	108.643650	0.442654	0.212715	2	0.027445	105.442245	14.380997	0	ISIC_0024309.jpg
4	0.015177	3.391087	2.449618	213.089690	0.480557	0.112513	2	0.037297	128.210530	15.006118	0	ISIC_0024311.jpg
...
2207	0.018336	3.556938	0.743756	53.281723	0.625389	0.097512	2	0.008995	83.130760	17.398417	1	ISIC_0034289.jpg
2208	0.011279	4.460854	0.480253	54.834350	0.442724	0.076281	3	0.004417	89.284060	20.241245	1	ISIC_0034294.jpg
2209	0.016355	4.947753	1.241271	99.550080	0.312866	0.148372	2	0.015413	111.825080	21.257566	1	ISIC_0034313.jpg
2210	0.019124	3.593563	3.074964	219.439000	0.384251	0.167293	4	0.065321	93.648860	24.786581	1	ISIC_0034316.jpg
2211	0.017508	3.037290	1.571407	116.866270	0.689444	0.051828	3	0.022832	74.727910	27.240215	1	ISIC_0034317.jpg

Fig. 6. 1: Dataset after feature extraction.

Next these ranked features have been cleaved into chunks from 6 to 10 and SVM has been applied on these feature categories of both 80% and 90% training of the data.

Fig. 6.3 illustrates the accuracies of different feature chunks using SVM on 80% training and 20% testing. With the chi-square value, the best accuracy of 75.17% has been achieved by both the top 9 and the top 10 features chunk.

However, on 90% training and 10% testing of the data, the highest accuracy of 77.93% has been achieved by the 10 features chunk which has been ornamented in Fig 6.4.



Fig. 6. 2: Feature ranking using Chi-Square.

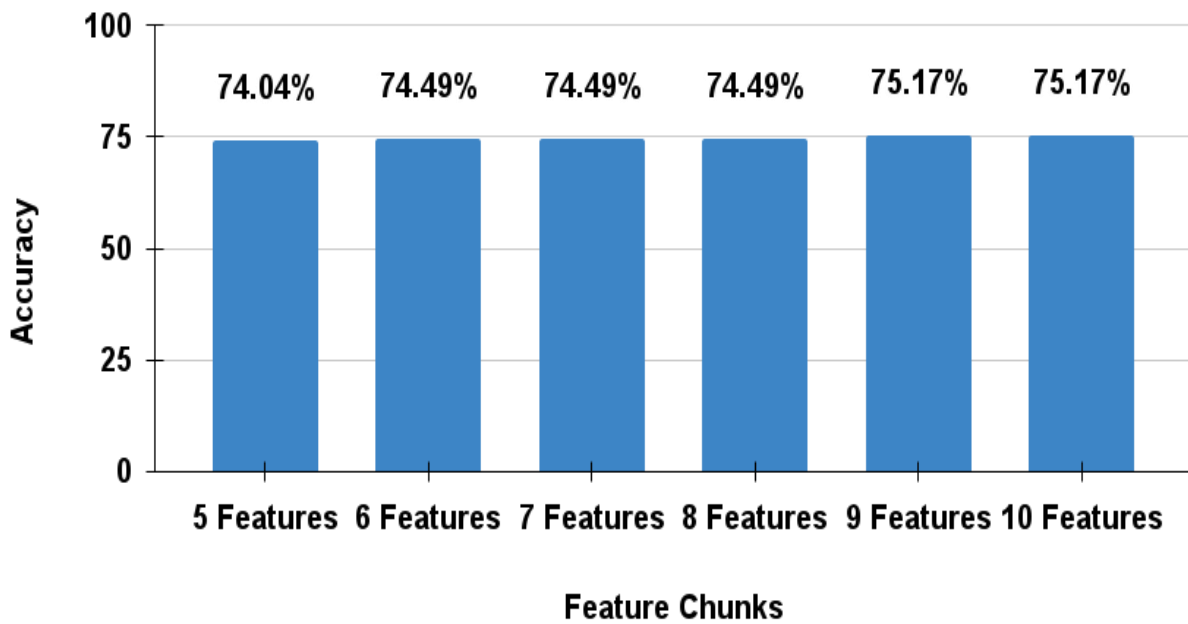


Fig. 6. 3: Accuracies of features chunks using SVM on 80% training and 20% testing.

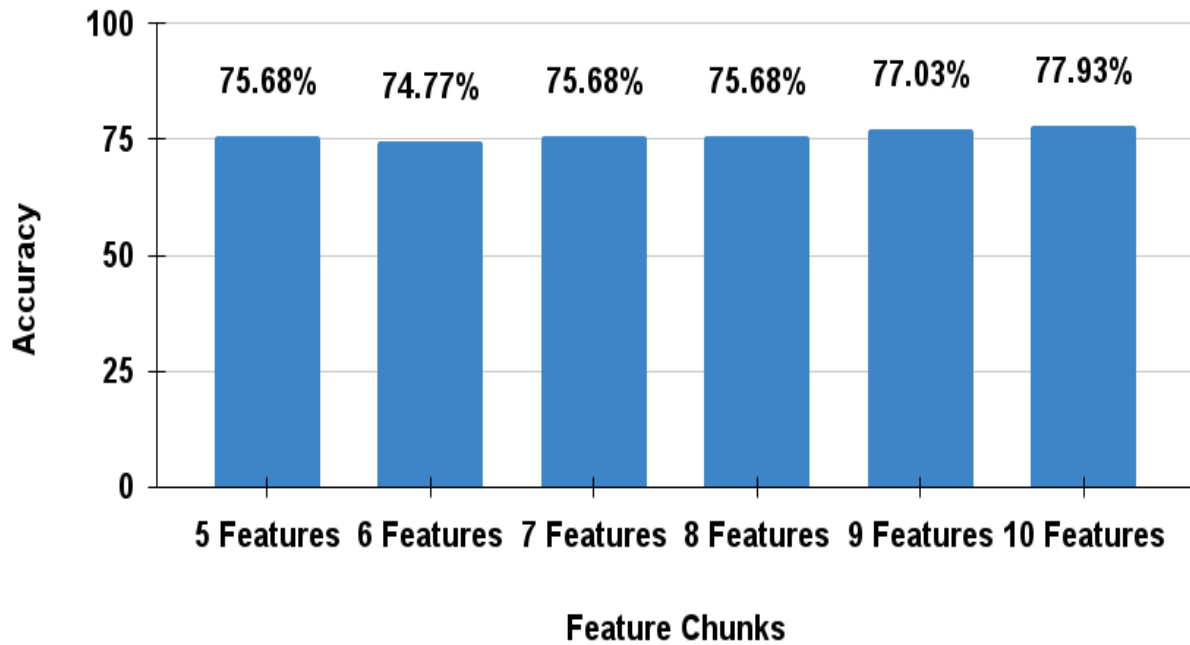


Fig. 6. 4: Accuracies of features chunks using SVM of 90% training and 10% testing.

We have advocated the correlation between all the features as well as with the target feature in this regard. The higher the correlation value among the feature indicates that one of the variables can be cleaved. Again the higher correlation between a variable and target feature reveals the importance of that feature on the target feature.

Figure 6.5 illustrates the correlation between the features as well as the target feature 'label'. Since the 'irc' and the 'ird' feature has gained the highest correlated value of 0.89, thus we have discounted the 'ird' feature since in terms of the target feature 'label', among the 'irc' and 'ird' feature, 'irc' has the upper hand with higher correlation value with 'label'. Again 'irc' and 'edge' has got correlation value of 0.81. Here we have discounted 'irc' since 'edge' has the upper hand with target feature 'label'.

After discounting 'irc', the correlation of the features with target feature has been ranked. Fig. 6.6 shows the feature ranking with target feature.

Next, with this rank we have cleaved these variables into chunks of 5 to 8 features to measure the accuracy with SVM. Fig. 6.7 shows the accuracies of different feature chunks on 80% training and 20% testing. It gives an insight that the top 8 features have achieved the highest accuracy of 75.17%.

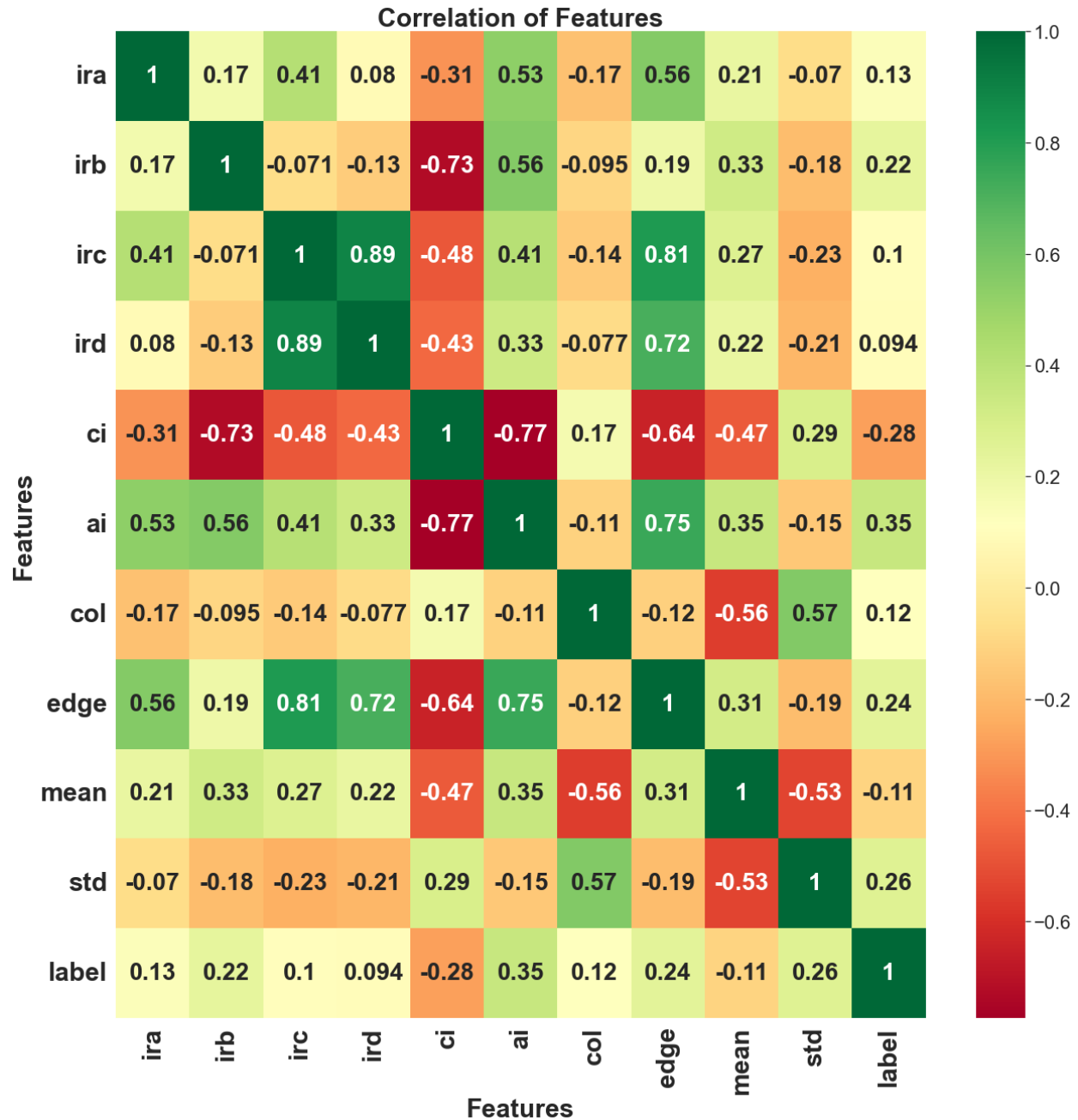


Fig. 6. 5: Correlation of features with target data.

Again Fig. 6.8 ornaments the accuracies of the feature chunks on 90% training, 10% testing. It achieves the highest accuracy of 76.58% with top 8 features.



Fig. 6. 6: Feature ranking after discounting 'irc' and 'ird'.

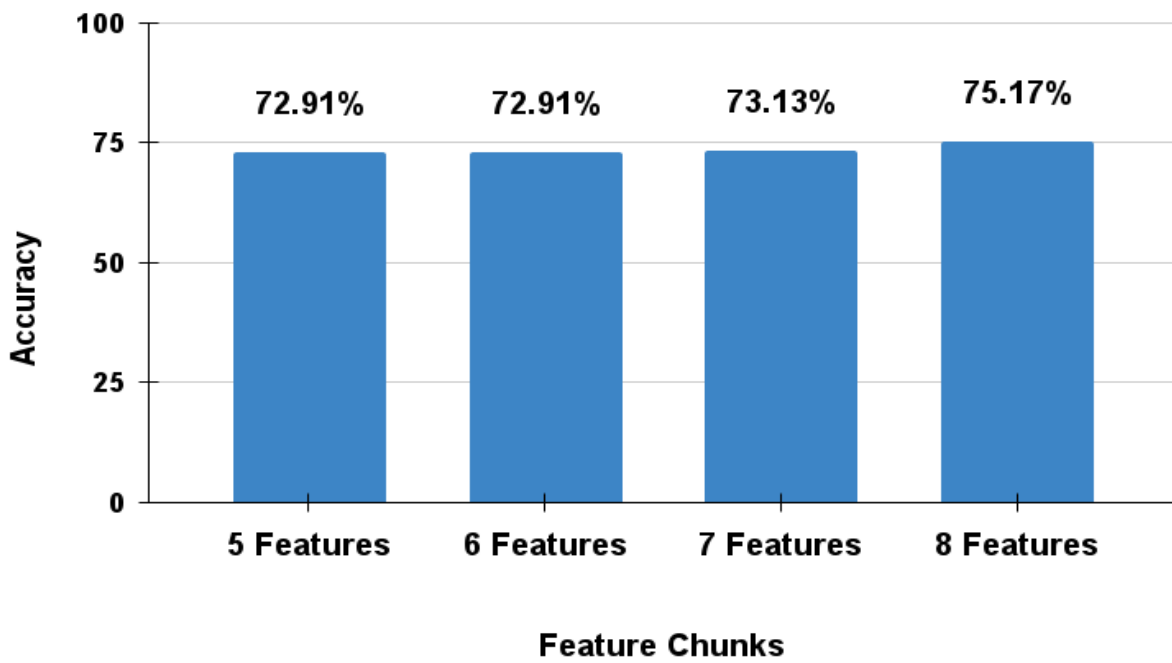


Fig. 6. 7: Accuracy of the features chunks in the correlation analysis using SVM on 80% training, 20% testing.

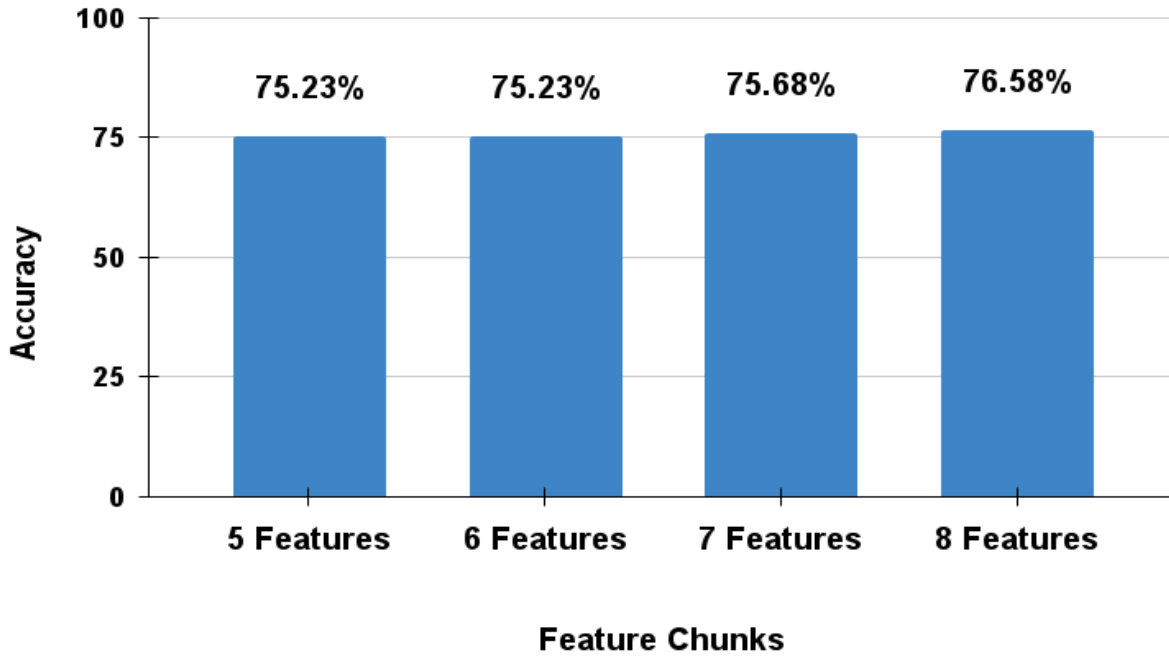


Fig. 6. 8: Accuracy of the features chunks in the correlation analysis using SVM on 90% training, 10% testing.

After the analysis of chi-square and correlation, it is visible that the best accuracy on both the 80% training, 20% testing, and the 90% training, 10% testing has been achieved by using all the 10 features altogether. Next the learning parameters of SVM are merged with the android system for prediction that has been ornamented in Chapter 7.

6.3.2 Accuracy

The testing accuracy of 80% training and 20% testing of the model is 75.17% whereas the testing accuracy for 90% training and 20% testing is 77.93%. Tables 6.2 and 6.3 show the confusion matrices of 80% training, 20% testing, and 90% training, 20% testing respectively.

Table 6. 2: Confusion matrix (80% Testing training, 20% testing)

	Positive = Malignant	Negative = Benign
Positive = Malignant	TP = 154	FN = 62
Negative = Benign	FP = 48	TN = 179

Table 6. 3: Confusion matrix (90% Training, 10% testing)

	Positive = Malignant	Negative = Benign
Positive = Malignant	TP = 74	FN = 27
Negative = Benign	FP = 22	TN = 99

6.3.3 Precision-Recall-Specificity-F1-Score

Table 6.4 shows the Precision-Recall-Specificity-F1-Score of both the 80% training, 20% testing, and 90% training, 10% testing.

Table 6. 4: Precision-Recall-Specificity-F1-Score of the test data.

Train	Test	Precision	Recall	Specificity	F1-Score
80%	20%	76.24%	71.30%	78.85%	73.68%
90%	10%	77.08%	73.27%	81.82%	75.13%

Figures 6.9 and 6.10 show the accuracy, precision, recall, specificity, and F1-Score of both the 80% training, 20% testing, and the 90% training, 10% testing respectively.

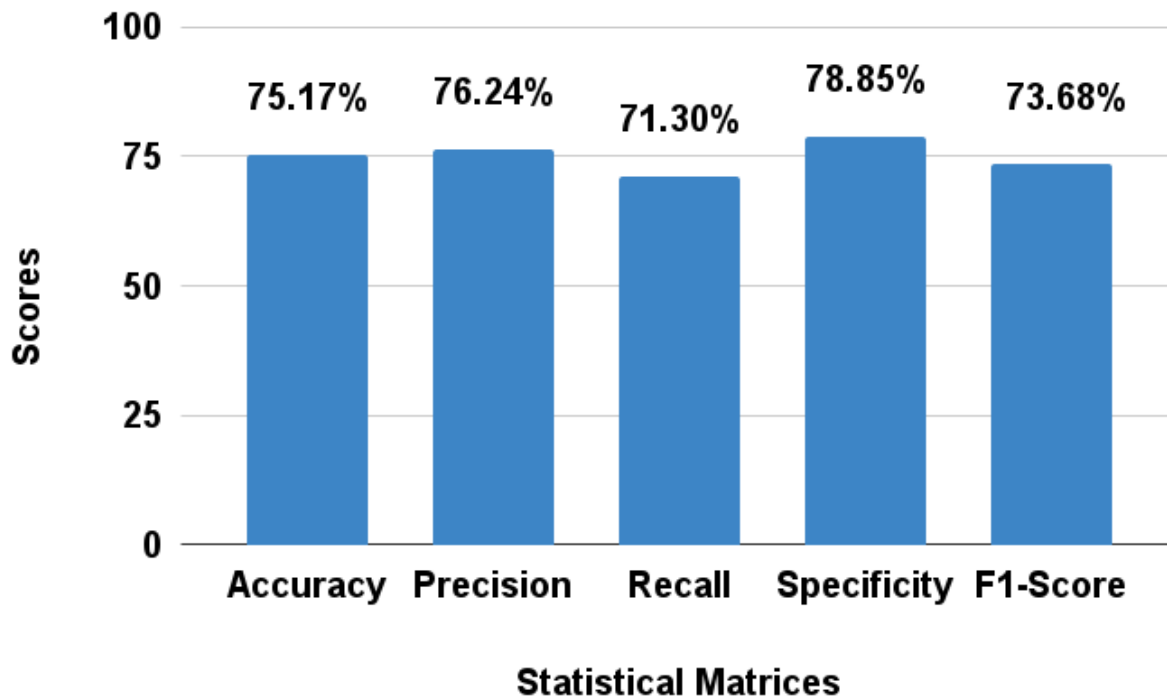


Fig. 6. 9: Performance measures of SVM on 80% training, 20% testing.

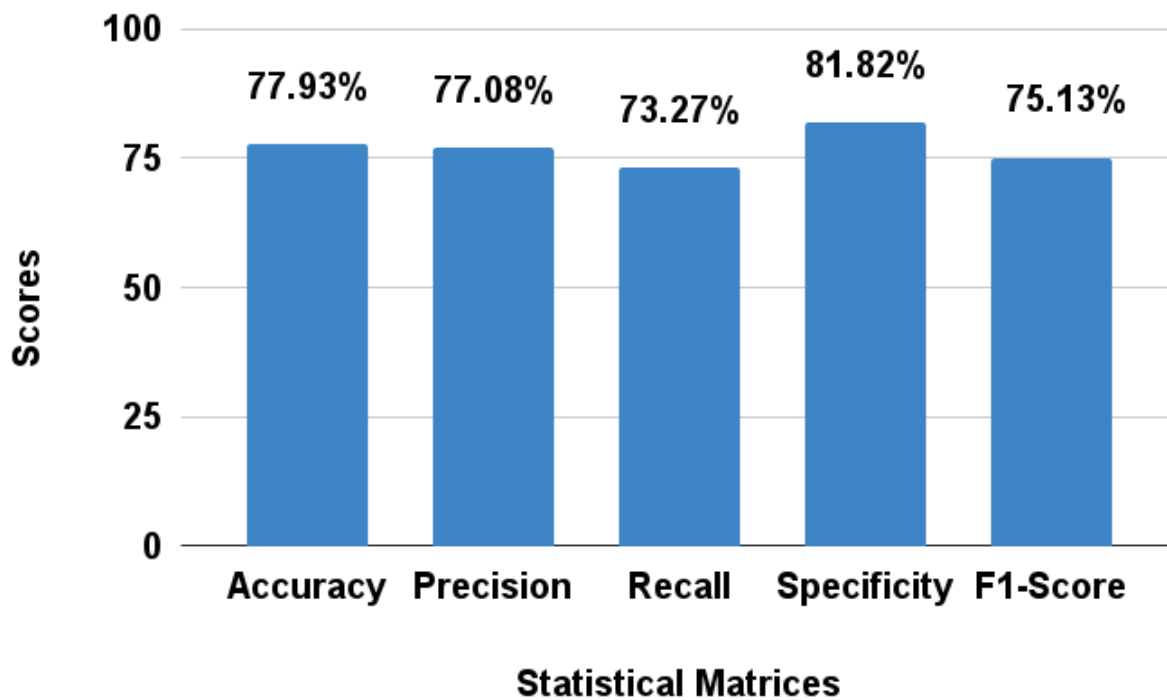


Fig. 6. 10: Performance measures of SVM on 90% training, 10% testing.

6.3.4 ROC Curve

Fig 6.11 and 6.12 show the ROC Curve of both the 80% training, 20% testing, and the 90% training, 10% testing respectively. The AUC of both the chunks are 0.75 and 0.78 respectively.

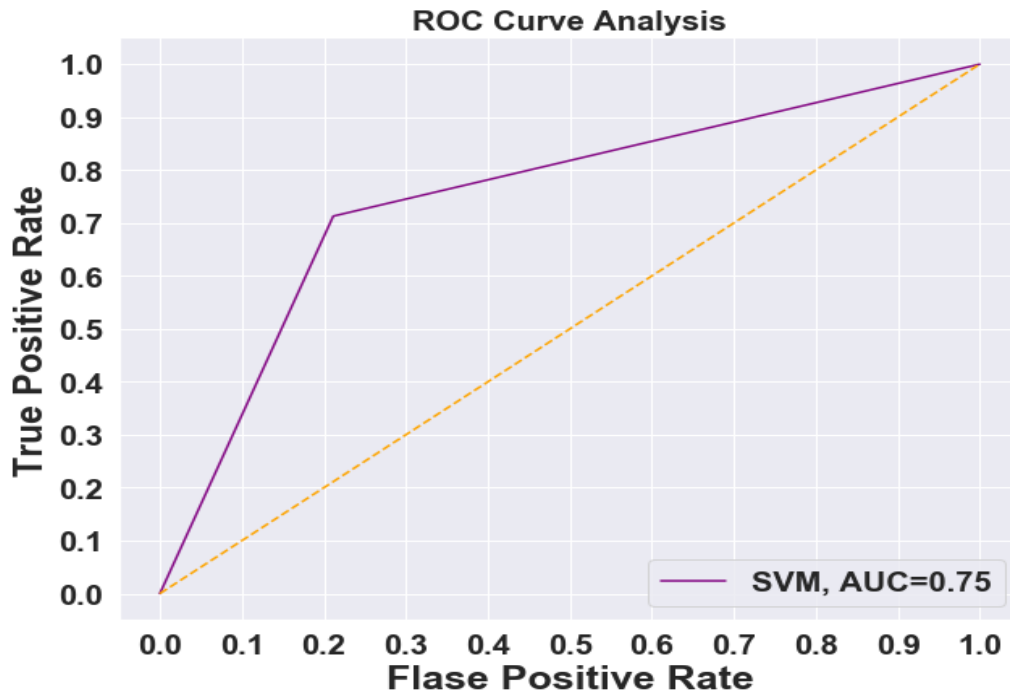


Fig. 6. 11: ROC Curve of SVM on 80% training, 20% testing data.

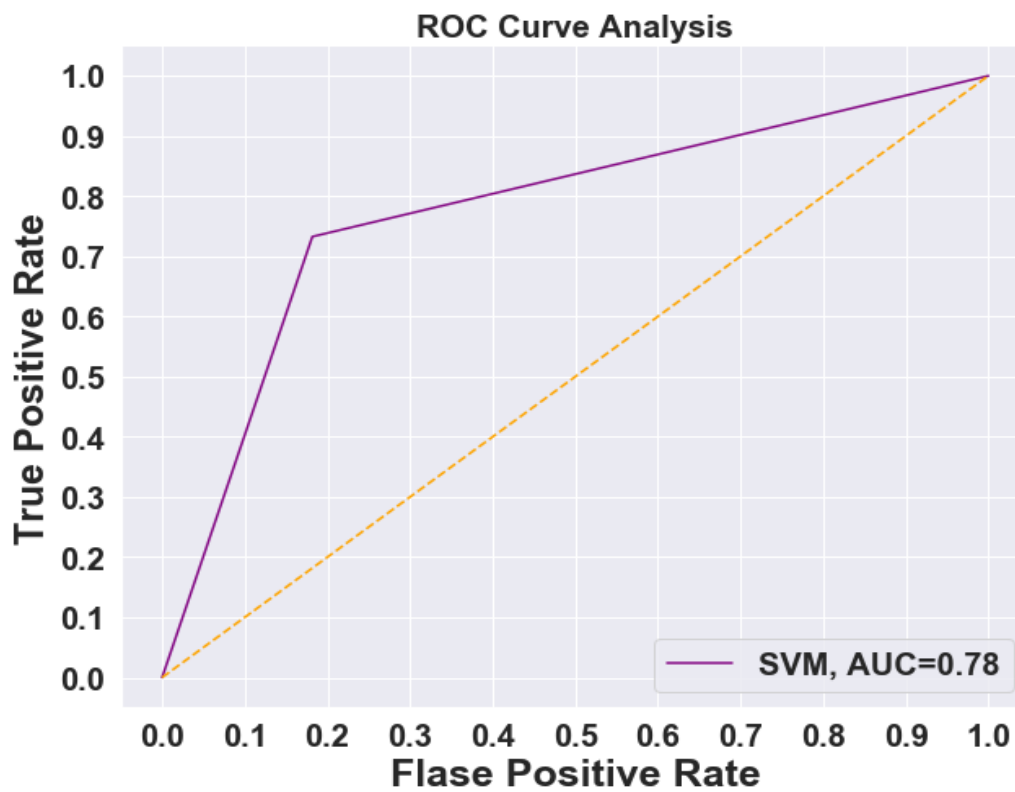


Fig. 6. 12: ROC Curve of SVM on 90% training, 10% testing data.

6.4 Comparison

The accuracy of our proposed methodology on the HAM10000 dataset is 75.17% for 80% training, 20% testing, and 77.93% for 90% training, 10% testing data. The comparison with different methods is listed in Table 6.5. The accuracy of Mporas et al. [22] on the HAM10000 dataset is 67.51%. Again, the accuracy of More et al. [23] on 90% training and 10% testing on the same dataset is 75%. Moreover, the accuracy on the same dataset with 10% testing of Pham et al. [26] is 74.75% and Purnama et al. [28] is 72%. However, Pai et al. [27] have achieved 78% accuracy on the same dataset with 20% test data and Garg et al. [24] had achieved 65.86% accuracy for the same. Mporas et al. [25] had achieved 74.69% accuracy on the same dataset.

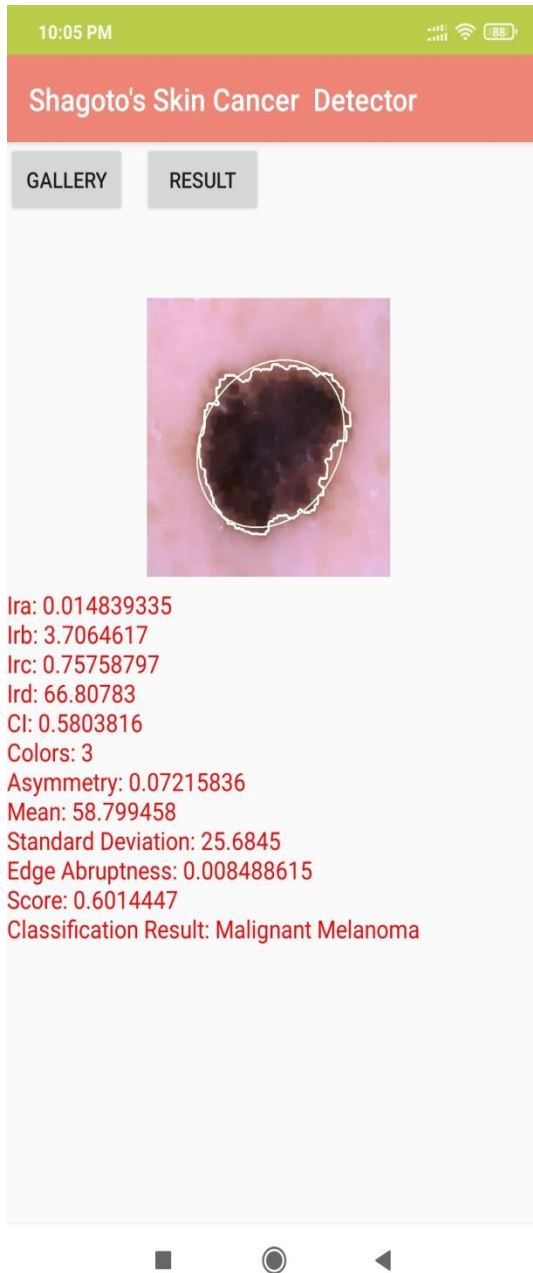
Table 6. 5: Comparison with other methods.

Year	Method	Dataset	Training Dataset	Testing Dataset	Accuracy
2020	Mporas et al. [22]	HAM10000	NA	NA	67.51%
2020	More et al. [23]	HAM10000	90%	10%	75%
2020	Garg et al. [24]	HAM10000	80%	20%	65.86%
2019	Mporas et al. [25]	HAM10000	NA	NA	74.69%
2019	Pham et al. [26]	HAM10000	90%	10%	74.75%
2019	Pai et al. [27]	HAM10000	80%	20%	78%
2019	Purnama et al. [28]	HAM10000	90%	10%	72%
2021	Proposed method	HAM10000	80%	20%	75.17%
2021	Proposed method	HAM10000	90%	10%	77.93%

Chapter 7

Implementation

The entire methodology is based on smartphone devices. To engender the process in smartphones, android studio [29] has been enabled. Image processing has been the cornerstone of this analysis. Image processing-related stuff have been handled by the OpenCV library [30] for android studio. After preprocessing, the dataset has been progressed to SVM for learning. For analyzing the learning with SVM, the Scikit-learn library [31] has been enabled in the python platform. For various visualizations, the Matplotlib library [32] and seaborn library [33] have been enabled. The parameters from the SVM has been added in the android studio for prediction. Fig. 7.1 shows the detection of malignant melanoma in two different smartphones: Xiaomi Note 6 Pro and Samsung Galaxy M10. Again Fig. 7.2 illustrates the detection of benign melanoma in Xiaomi and Samsung mobiles. Both the figures allude to the identical values in different smartphones on the same lesion image.

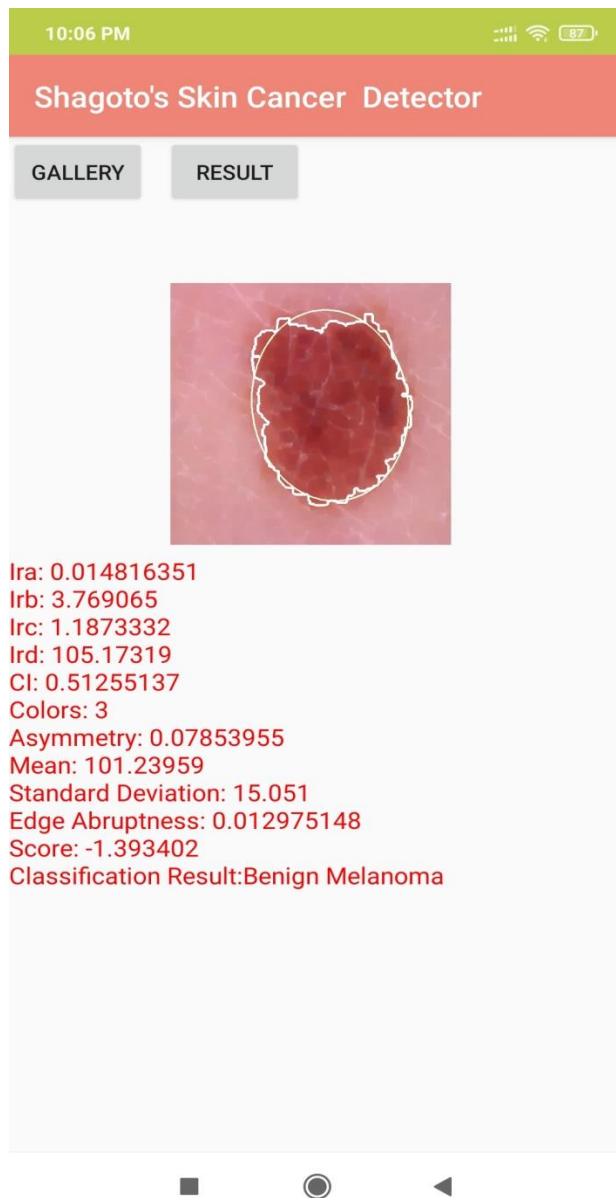


(a) Xiaomi Redmi Note 6 Pro



(b) Samsung Galaxy M10

Fig. 7. 1: Detection of malignant melanoma in two different smartphones.



(a) Xiaomi Redmi Note 6 Pro



(b) Samsung Galaxy M10

Fig. 7. 2: Detection of benign melanoma in two different smartphones.

Chapter 8

Conclusion

8.1 Conclusion

Recognizing the severity of skin cancer these days, our proposed method works at the detection of skin cancer using skin lesion images in smartphones. The accuracy of our proposed methodology is 75.17% for 80% training, 20% test data, and 77.93% for 90% training, 10% test data.

Our experimental results signify that our proposed method has a comparable execution to other methods taking the same percentage of training and testing data proportions. In near future, we have intentions to work with different datasets with various algorithms in terms of learning.

8.2 Future Work Direction

We are looking forward to attaching ensemble models and also adding neural network algorithms with various methods apart from ABCD rules in the future.

Bibliography

- [1] A. Kopf, T. Salopek, J. Slade, A. Marghoob and R. Bart, "Techniques of cutaneous examination for the detection of skin cancer", *Cancer*, vol. 75, no. 2, pp. 684-690, 1995.
- [2] "Skin Cancer Facts & Statistics," [Online]. Available: <https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/>. [Accessed: 02- Jan- 2022].
- [3] M. Weinstock et al., "Skin biopsy utilization and melanoma incidence among Medicare beneficiaries", *British Journal of Dermatology*, vol. 176, no. 4, pp. 949-954, 2017.
- [4] M. Giger and K. Suzuki, "Computer-Aided Diagnosis", *Biomedical Information Technology*, p. 359-XXII, 2008.
- [5] F. Nachbar et al., "The ABCD rule of dermoscopy", *Journal of the American Academy of Dermatology*, vol. 30, no. 4, pp. 551-559, 1994. Available: 10.1016/s0190-9622(94)70061-3 [Accessed 2 January 2022].
- [6] H. Firmansyah, E. Kusumaningtyas and F. Hardiansyah, "Detection melanoma cancer using ABCD rule based on mobile device", in *2017 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*, Surabaya, Indonesia, 2017, pp. 127-131.
- [7] M. Hearst, S. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines", *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18-28, 1998.
- [8] S. Alizadeh and A. Mahloojifar, "A Mobile Application for Early Detection of Melanoma by Image Processing Algorithms", *2018 25th National and 3rd International Iranian Conference on Biomedical Engineering (ICBME)*, Qom, Iran, 2018.
- [9] S. Shafiq, P. Prasad, A. Alsadoon, S. Ali and A. Elchouemi, "Computer Aided Early Detection and Classification of Malignant Melanoma", *2018 10th International Conference on Computational Intelligence and Communication Networks (CICN)*, Esbjerg, Denmark, 2018.
- [10] DermNet NZ – All about the skin | DermNet NZ", *Dermnetnz.org*, 2022. [Online]. Available: <https://dermnetnz.org/>. [Accessed: 02- Jan- 2022].
- [11] ISIC Archive. [Online]. Available: <https://www.isic-archive.com/#!/topWithHeader/onlyHeaderTop/gallery?filter=%5B%5D>. [Accessed: 02- Jan- 2022].

- [12] M. Taufiq, N. Hameed, A. Anjum and F. Hameed, "m-Skin Doctor: A Mobile Enabled System for Early Melanoma Skin Cancer Detection Using Support Vector Machine", Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp. 468-475, 2016.
- [13] P. Tschandl, C. Rosendahl and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions", Scientific Data, vol. 5, no. 1, 2018.
- [14] G. Deng and L. Cahill, "An adaptive Gaussian filter for noise reduction and edge detection", 1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference, San Francisco, CA, USA, 1993, pp. 1615-1619 vol.3.
- [15] Z.Fang, M.Yulei, Z.Junpeng, "Medical Image Processing Based on Mathematical Morphology", The 2nd International Conference on Computer Application and System Modeling (2012).
- [16] L. Jianzhuang, L. Wenqing and T. Yupeng, "Automatic thresholding of gray-level pictures using two-dimension Otsu method", in China., 1991 International Conference on Circuits and Systems, Shenzhen, China, 1991.
- [17] T. Kanimozhi and A. Murthi, "COMPUTER AIDED MELANOMA SKIN CANCER DETECTION USING ARTIFICIAL NEURAL NETWORK CLASSIFIER", Singaporean Journal of Scientific Research(SJSR), vol. 8, no. 2, pp. 35-42, 2016..
- [18] N. K. EL Abbadi and Z. Faisal, "Detection and Analysis of Skin Cancer from Skin Lesions", International Journal of Applied Engineering Research, vol. 12, no. 19, pp. 9046-9052, 2017.
- [19] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, 1995, pp. 338-345: Morgan Kaufmann Publishers Inc.
- [20] Patel, Harsh & Prajapati, Purvi. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering. 6. 74-78.
- [21] Y. Tian, Y. Shi and X. Liu, "RECENT ADVANCES ON SUPPORT VECTOR MACHINES RESEARCH", Technological and Economic Development of Economy, vol. 18, no. 1, pp. 5-33, 2012.

- [22] I. Mporas, I. Perikos and M. Paraskevas, "Color Models for Skin Lesion Classification from Dermatoscopic Images", *Advances in Integrations of Intelligent Methods*, pp. 85-98, 2020.
- [23] J. More, M. Nath, P. Yamgar and A. Bhatt, "Skin Disease Classification Using Convolutional Neural Network", *International Research Journal of Engineering and Technology (IRJET)*, vol. 07, no. 05, pp. 7644-7647, 2020.
- [24] R. Garg, S. Maheshwari and A. Shukla, "Decision Support System for Detection and Classification of Skin Cancer Using CNN", *Advances in Intelligent Systems and Computing*, pp. 578-586, 2020.
- [25] I. Mporas, I. Perikos and M. Paraskevas, "Pigmented Skin Lesions Classification Using Data Driven Subsets of Image Features", 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), Patras, Greece, 2019.
- [26] T. Pham, G. Tran, T. Nghiem, A. Doucet, C. Luong and V. Hoang, "A Comparative Study for Classification of Skin Cancer", 2019 International Conference on System Science and Engineering (ICSSE), Dong Hoi, Vietnam, 2019.
- [27] K. Pai and A. Giridharan, "Convolutional Neural Networks for classifying skin lesions", *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, Kochi, India, 2019.
- [28] I. Purnama et al., "Disease Classification based on Dermoscopic Skin Images Using Convolutional Neural Network in Teledermatology System", 2019 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), Surabaya, Indonesia, 2019.
- [29] "Android Developers", Android Developers, 2022. [Online]. Available: <https://developer.android.com/>. [Accessed: 02- Jan- 2022].
- [30] "Android - OpenCV", OpenCV, 2022. [Online]. Available: <https://opencv.org/android/>. [Accessed: 02- Jan- 2022].
- [31] "sklearn.svm.LinearSVC", scikit-learn, 2022. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>. [Accessed: 02- Jan- 2022].
- [32] "Matplotlib — Visualization with Python", Matplotlib.org, 2022. [Online]. Available: <https://matplotlib.org/>. [Accessed: 02- Jan- 2022].
- [33] "seaborn: statistical data visualization — seaborn 0.11.2 documentation", Seaborn.pydata.org, 2022. [Online]. Available: <https://seaborn.pydata.org/>. [Accessed: 02-

Jan- 2022].