Early detection of malignant melanoma using random forest algorithm

S. Amutha¹
¹National engineering college, ECE, amuthashenbagaraj@email.com

Mrs. R. Manjula Devi²
²National engineering college, ECE, manju.rengasamy@email.com

Abstract—Melanoma is a type of skin cancer, which is caused by the development of the malignant tissues of the melanocytes. Majority of deaths caused by the skin cancer are as a result of melanoma that is inferred from the statistical evidence. This paper introduces the skin lesion analysis system for the early detection of melanoma. The skin lesion analysis system aims to focus the efficiency of the classification results. The random forest algorithm is proposed in lesion classification process, which enhances the importance of the predictor's variables and a measure of the internal structure of the data. The experimental results show that the proposed system is efficient, achieving classification of the benign, atypical and melanoma images with the accuracy.

Index Terms—Random forest, Skin cancer, Melanoma.

1 INTRODUCTION

Melanoma is the most dangerous type of skin cancer. Melanoma also known as malignant melanoma is a type of cancer that developed from the pigment containing cells known as melanocytes. Melanoma can occur on any skin surface. In men, it's often found on the skin on the head, on the neck, or between the shoulders and the hips. In women, it's often found on the skin on the lower legs or between the shoulders and the hips. Melanoma is rare in people with dark skin. When it does develop in people with dark skin, it’s usually found under the fingernails, under the toenails, on the palms of the hands, or on the soles of the feet. Ultraviolet (UV) rays are clearly a major cause of UV rays can damage the DNA in skin cells. Sometimes this damage affects certain genes that control how skin cells grow and divide. If these genes no longer work properly, the affected cells may form a cancer. Sunlight is the main source of UV rays. The nature of the UV exposure may play a vital role in melanoma development.

Skin cancer is the most commonly diagnosed cancer in the United States, yet most cases are preventable. Every year in the United States, nearly 5 million people are treated for skin cancer, at an estimated cost of $8.1 billion. Melanoma, the deadliest form of skin cancer, causes more than 9,000 deaths each year. Unlike many other cancers, skin cancer rates have continued to rise in recent years. Here are the American Cancer Society's estimates for melanoma in the United States for 2015:

- About 73,870 new melanomas will be diagnosed (about 42,670 in men and 31,200 in women).
- About 9,940 people are expected to die of melanoma (about 6,640 men and 3,300 women).

Although survival rate is increasing, death rate from malignant melanoma is exponentially increasing. Early diagnosis is crucial for the treatment, because malignant melanoma is very invasive when it affects melanocyte.

2 RELATED WORK

In dermatology, there has been a long interest in exploiting computer technology. Recent years have seen increased activities in developing in machine learning and computer vision techniques for skin lesion diagnosis, especially for diagnosing melanoma case. Braun et al. have worked on [1] dermoscopy research tool and also cover different aspects, such as the new equipment, new structures, the importance of blood vessels, etc. There were several drawbacks in dermoscopy; it could not be used for clinically suspicious skin lesions.

Silveira et al. proposed the early diagnosis of malignant melanoma, but their interpretation is time consuming and subjective, even for trained dermatologists[2]. Six different segmentation methods are Adaptive thresholding, Gradient vector flow, Adaptive snake, Level set method of Chan et al., Expectation-maximization level set, Fuzzy-based split-and-merge algorithm were compared and evaluated by four metrics (HM, TDR, FDR, HD). Out of six segmentation methods, only AS and EM-LS methods are robust and useful for the lesion segmentation to assist the clinical diagnosis of dermatologists.

Rademaker et al. [3] introduced digital monitoring by whole body photography and sequential digital dermoscopy detects thinner melanomas. Patients undergoing whole-body photography and sequential digital dermoscopy are largely self-referred. Melanoma is not detected until they are quite advanced.

Abbas and Celebic introduced a comparative study of the state-of-the-art hair-repaired methods with a novel algorithm is also proposed by morphological and fast marching schemes [4]. Non-linear partial differential equation based methods are not texture-based inpainting methods and, it was not suitable for hair removal in dermoscopic images.

Suer et al. [5] developed a automated assessment tools for dermoscopy images have become an important research field mainly because of inter- and intra-observer variations in human interpretation. It works on color image without preprocessing and it cannot find any point density-reachable from the starting point. This
procedure followed until all of the points in the EPS neighborhood are touched or visited at least once.

Wadhawan et al. introduced a portable library for melanoma detection on handheld devices [6]. The portable library for automated detection of melanoma termed SkinScan that can be used on smart phones and other handheld devices. User cannot use this application on other Portable device. This application is only suitable for Apple iphone4.

Doukas et al. developed a system consisting of a mobile application that could obtain and recognize moles in skin images and categorize them according to their brutality into melanoma, nevus, and benign lesions. As indicated by the conducted tests, Support Vector Machine (SVM) resulted in only 77.06% classification accuracy [7].

Garcia et al. proposed a automatically detects and removes hairs and ruler markings from dermoscopy images [8]. This method addresses the issue of preserving morphological features during artifact removal. It has been adopted by the Dull Razor hair-removal software. However, this global approach can sometimes lead to unsatisfactory results.

Abuzaghleh et al. introduced [9] an image recognition technique, user will be able to capture skin images of different mole types. System will analyze and process the images and alert the user at real-time to seek medical help urgently. It makes a significant impact on health care delivery as assistive devices in underserved and remote areas. System couldn’t allow the user to captured images using the smart phone.

From the above literature review it is clear that the segmentation is achieved by means of morphological operation. The lesion segmentation has impact of unwanted blur on the result images. The improvement is as achieved by several techniques.

### 3 Proposed Methodology

<table>
<thead>
<tr>
<th>Image acquisition</th>
<th>Hair detection &amp; Exclusion</th>
<th>Lesion segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Block diagram for the proposed dermoscopy image analysis system.

Proposed work uses graph partitioning method are an effective mechanism for image segmentation since they model the impact of pixel neighborhoods on a given cluster of pixels or pixel, under the assumption of homogeneity in images.

#### 3.1 Image Acquisition

The action of retrieving an image from some source is usually a hardware-based source. Performing image acquisition in image processing is always the first step in the workflow sequence because, without an image, no processing is possible. The size of the captured lesions will vary based on the distance between the capturing the images in different light environment will be another challenge of the lesion will not be clearly visible. To overcome these challenges, a dermoscope are used to capture high quality images, here data set images are used.

#### 3.2 Hair Detection & Exclusion

Removing the hairs without altering the lesion is important. Clean hair mask is to segment and remove the hair in the image, preparing it for further segmentation and analysis. To detect and exclude the hairs, bottom hat filters are used afterwards kernel hair mask can be used.

After the hair mask, the image is reconstructed to fill the hair gap with actual pixels.

#### 3.3 Filtering

Filtering is a technique for modifying or enhancing an image. Gaussian and median filters are used. The Gaussian filter is a non-uniform low pass filter. A kernel is a small matrix of numbers that is used in image convolutions. The kernel coefficients diminish with increasing distance from the kernel’s centre. Central pixels have a higher weighting than those on the periphery. The Standard deviation of the Gaussian function plays an important role in its behavior. Larger values of $\sigma$ produce a wider peak. 2D Gaussian low pass filters is generated by,

$$h_g(n_1, n_2) = e^{-\frac{n_1 + n_2}{2\sigma^2}}$$

(1)

$$h(n_1, n_2) = h_g(n_1, n_2)$$

(2)

Where,

- $h$ is a 2-D filter of size $n_1, n_2$
- $\sigma = 0.5$ (The default value for sigma is 0.5)

Median filter is a popular low-pass filter, attempting to remove noisy pixels while keeping the edge intact. However, the median is a more robust average than the mean and so a single very unrepresentative pixel in a neighborhood will not affect the median value significantly. The values of the pixel in the window are stored and the middle value in the sorted list (or average of the middle two if the list has an even number of elements) is the one plotted into the output image. The median filtered image $g(x, y)$ can be obtained from the median pixel values in a neighborhood of $(x, y)$ in the input image $f(x, y)$, as defined by the following formula,

$$g(x, y) = median \left[ \sum_{i=1}^{1} \sum_{j=1}^{1} f(x-i, y-j) \right]$$

(3)

Median filters are great at preserving edges and eliminating impulse noise.
3.4 IMAGE SEGMENTATION

Pigmented skin lesion segmentation is used to separate the lesion from the background; it is an essential process before starting with the feature extraction in order to classify the three different types of lesion. After the Gaussian filter is applied, a global threshold is computed by Otsu’s method to convert an intensity image to a binary image. Otsu’s method chooses the threshold to minimize the intra-class variance of the background and foreground pixels.

Next, the graph cut method is applied. It is an effective tool for image segmentation since they model the impact of pixel neighborhoods on a given cluster of pixels or pixel, under the assumption of homogeneity in images. In these methods, the image is modeled as a weighted, undirected graph. Usually a pixel or a group of pixels are associated with nodes and edge weights define the (dis)similarity between the neighborhood pixels.

The graph (image) is then partitioned according to a criterion designed to model "good” clusters. Each partition of the nodes (pixels) output from these algorithms are considered an object segment in the image. Under graph cut method max flow min theorem can be used. The max-flow min-cut theorem states that in a flow network, the maximum amount of flow passing from the source to the sink is equal to the minimum capacity that, when removed in a specific way from the network, causes the situation that no flow can pass from the source to the sink.

3.5 FEATURE EXTRACTION

When the input data to an algorithm is too large to be processed and it is suspected to be redundant then it can be transformed into a reduced set of features. The extracted features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data. There four features are calculated namely Lesion shape, Orientation, Margin, Intensity pattern feature.

The feature set includes five parameters that are calculated from the detected pigment network as follows,

1. Pigment network vs. Lesion area ratio (f1)

   The ratio between the area of the detected pigment network
   \( A(P) \) and the area of the segmented lesion \( A(L) \).
   \[ f_1 = \frac{A(P)}{A(L)} \]  

2. Pigment network area vs. Filled network area network (f2)

   The ratio between the pigment network area \( A(P) \) and the
   filled network area \( A(F) \).
   \[ f_2 = \frac{A(P)}{A(F)} \]  

3. Total number holes in the pigment in the network(f3)

   \[ f_3 = \sum H \]  

   Where,
   \( H \) denotes the hole in the pigment network

4. Total number of holes in the pigment network vs. Lesion area ratio(f4)

   \[ f_4 = \frac{\sum H}{A(L)} \]  

5. Total number of holes in the pigment network vs. Filled network area ratio(f5)

   \[ f_5 = \frac{\sum H}{A(F)} \]  

3.6 CLASSIFICATION

Classification is a general process related to categorization, the
process in which ideas and objects are recognized, differentiated, and
understood. There are three different classifiers are used. Classifier is
proposed to classify the image into three categories (i.e., Benign, Atypical and Melanoma). Random forest (RF) classifier is used. The
RF consists of a number of trees, with each tree grown using some
form of randomization and other tasks that operate by constructing a
multitude of decision trees at training time and outputting the class
that is the mode of the classes or mean prediction of the individual
trees. The random forests algorithm (for both classification and
regression) is as follows:

1. Create \( n_{tree} \) bootstrap samples from the original data.

2. For each of the bootstrap samples, grow an unpruned classification or regression tree with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample \( m_{try} \) of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when \( m_{try} = p \), the number of predictors.)

3. Predict new data by aggregating the predictions of the
   \( n_{tree} \) trees (i.e., majority votes for classification, average
   for regression).

An estimate of the error rate can be obtained, based on the training
data, by the following:

- At each bootstrap iteration, predict the data not in the bootstrap sample (“out-of-bag” or OOB) using the tree grown with the bootstrap sample.
- Aggregate the OOB predictions. Calculate the error rate, and call it the OOB estimate of error rate.

4 RESULT AND CONCLUSION

In this work, the dataset image are taken as a input image with the
hair particles and thus fed into a bottom hat filter in order to detect
the hair in an input image. The channels can be separated as R, G, B
and the hair will be removed as shown in Figure 4.
In filtering process median and Gaussian are used to remove the noise in an image. Median filtering is a nonlinear process useful in reducing salt-and-pepper noise and also preserves the edges. Gaussian smoothing is a linear process very effective for removing noise. The filtered images are shown in Figure 5.

Neighborhood pixels are not clearly segmented in the lesion segmentation process. To overcome this problem Graph Cut method is used. The graphical representation of graph cut method is shown in figure 6.

In the feature extraction different set of features are calculated. 2D Fast Fourier Transform (FFT), 2D Discrete Cosine Transform (DCT) outputs are shown in figure 7.

Receiver Operating Characteristic curve is a plot of the true positive rate against the false positive rate for the different possible cut points of a diagnostic test. Classifier distributions and recognized disease are shown in figure 10.

The classification results are compared between supporting vector machine and random forest. The error rate of random forest is less than the existing algorithm which is shown in figure 11.
Therefore, the random forest gives the better performance.

5 CONCLUSION

In this work, the skin lesion analysis system is addressed to detect the melanoma at early stage. The lesion classification process with random forest is proposed to enhance the efficiency of classification results. The results are shows that the proposed system is efficient, achieving classification of the benign, atypical and melanoma images with the accuracy. Future work would focus on early detection malignant melanoma using Extreme learning machine classification algorithm to reduce the error rate and further optimize the performance.

REFERENCES


AUTHOR PROFILE:
- Author S. Amutha is currently pursuing masters degree in electronics communication engineering in National engineering college, Kovilpatti, India. E-mail: amuthashenbagaraj@mail.com
- Co-Author Mrs. R. Manjula Devi is currently working as assistant professor in electronics communication engineering in National engineering college, Kovilpatti, India. E-mail: manju.rengasamy@mail.com