Skin Cancer Classification Using K-Means Clustering

1 Mohd Anas, 2 Ram Kailash Gupta, 3 Dr. Shafeeq Ahmad
1 M. Tech Student, Dept. of Computer Science & Engineering, AIET
2 Asst. Prof., Dept. of Computer Science & Engineering, AIET
3 Prof. & Director, Dept. of Computer Applications, AIET
Lucknow, Uttar Pradesh, India

Abstract—Detection of skin cancer gives the best chance of being diagnosed early. Biopsy method for skin cancer detection is much painful. Human interpretation contains difficulty and subjectivity therefore automated analysis of skin cancer affected images has become important. This paper proposes an automatic medical image classification method to classify two major type skin cancers: Melanoma, and Non-melanoma. In this paper, we have used the color and texture features in combination which gives better results than using color or gray level information alone. We have used k-means clustering algorithm to segment the lesion. The features are extracted by six different color-texture feature extractors from the segmented images. Classification accuracy of our proposed system is evaluated on four different types of classifiers and their values are compared with one another. The results of the proposed system are computed on five different classification rate in order to perform better analysis of our proposed system.

Keywords: K-means clustering, segmentation, local binary pattern and color percentile.

I. INTRODUCTION

Skin Cancer incidence is increasing at 3.1% per year [1]. Skin cancer spread over the body with the help of lymphatic and blood vessels. Thus, early detection of skin cancer is very important for proper diagnosis of the disease.

Melanoma and Non-Melanoma are two major categories of skin cancers. Malignant melanoma is of several sub-types. Basal cell carcinoma and Squamous cell carcinomas are two main types of non-melanoma skin cancers.

Each type of skin cancer is different from the other skin cancers in certain characteristics.

In clinical detection of skin cancer diagnosis, dermatologist use a visual inspection. Clinical diagnostic performance is very poor in comparison to dermoscopy and automatic diagnosis. Dermoscopy is a non-invasive diagnostic technique. It uses clinical dermatology and dermatopathology in combination to inspect the morphological features which is not possible in clinical detection. Dermoscopy increases the performance of diagnosis with 10-30% compared to unaided eye [2]. Differentiation of skin cancer images needs much more experience with dermoscopy technique. Less experienced clinicians use ABCD-E rule to improve the diagnostic performance [3], which are as follows:

A- Asymmetry: If you draw a line through the middle, the two sides will match, meaning it is symmetrical. If you draw a line through lesion, the two halves will not match, meaning it is asymmetrical.

B- Border: The borders of an early melanoma tend to be uneven.

C- Color: A number of different shades of brown, tan or black could appear. A melanoma may also become red, white or blue.

D- Diameter: Benign moles usually have a smaller diameter than malignant ones.

E- Evolving: Be on the alert when a mole starts to evolve or change in any way.

Automatic image processing of skin cancer gives better results by providing the exact information about lesion, which can be useful for the clinician to detect and classify skin cancer. It is also used as a standalone early warning tool. Effective implementation of this automatic technique may give reduced deaths with benefits both to the patients and to the health care system. Working of automatic skin cancer detection has three main stages: (1) Segmentation of lesion (2) Feature extraction and feature selection (3) Lesion classification.

Computer aided classification tools are important in medical imaging for diagnosis and evaluation. Predictive models are used in a variety of medical domains for diagnostic and prognostic tasks. These models are built based on experience which constitutes data acquired from actual cases. The data can be preprocessed and expressed in a set of rules, such as that it is often the case in knowledge-based expert systems, and consequently can serve as training data for statistical and machine learning models.

The general approach of developing a computer aided classification system for the diagnosis of skin cancer is to find the location of a lesion. The first step is to establish a standard general scheme of a computer aided classification system for skin lesions. The inputs to the computer aided system are digital images obtained by dermoscopic method, with the possibility to add other acquisition system such as ultrasound or confocal microscopy. In the first phase preprocessing of image is done that allows reducing the ill effects and various artifacts like hair that may be present in the dermoscopic images. It is followed by the detection of the lesion by image segmentation technique. Once the lesion is localized, different chromatic and morphological features can be quantified and used for classification. Differentiation of melanoma and non-
melanoma images demands very fast image processing and feature extraction and classification algorithms.

The rest of the paper is organized as follows: Section II present the motivation proposed algorithm and implementation. Section III present the performance evaluation. Section IV present the conclusion and direction for future enhancement.

II. MOTIVATION

The analysis of medical images has always been performed visually by physicians. In the last two decades, a strong impulse has been given to develop automated systems capable of assisting physicians in this task, mainly because they have the desire of quantifying a number of anatomical and functional parameters, useful for diagnosis and therapy, which can be evaluated only qualitatively by human beings. Presence of noise and of masking structures, the variability of biological shapes and tissues, imaging system anisotropy etc. make the automated analysis of medical images a very hard task.

Segmentation is an important process in image processing applications and computer vision because doctors are interested in certain regions of the dermoscopy image. Segmentation divides an image into a number of discrete regions. Pixels, in each region, have high similarity and high contrast between regions. Many Researchers use only gray-level for image segmentation [3]. But, in our proposed system, we use color information of the image for lesion segmentation.

In general, we covert the color image in gray-level image therefore color information does not used. There is a wide variety of segmentation methods used in dermoscopy images [4]. Recent advancements include thresholding [5], k-means clustering [6], fuzzy c-means clustering [7], density-based clustering, mean shift clustering , gradient vector flow snakes , color quantization followed by spatial segmentation , statistical region merging, watershed transformation, dynamic programming , and supervised learning. Clustering is an unsupervised learning technique, where we give the number of clusters in advance to classify pixels. A similarity measure is defined between pixels and similar pixels are then grouped into a clusters. We use k-means clustering for segmentation of color images.

It is very hard to differentiate skin cancer visually. Identification and extraction of most effective features from cancer affected lesion is very important. Each class of skin cancer has some different features than others. We use these different features for classification.

Texture is important aspect of an image that has been widely used in medical image classification, visual inspection, remote sensing. Textures are generally complex visual patterns that have characteristics such as brightness, color, slope, and size. Texture features can be extracted in several methods, namely, statistical, structural, model-based, and transform information. Feature extraction extracts useful features or properties from original dataset of an image. These extracted features easily classify the classes of skin cancer.

Color features are mainly statistical parameters. These are calculated from inter and intra-channel of an image, like average value and standard deviation of the RGB [3] or HSV color channels [4]. Here, we use “Local Binary Patterns + Color percentiles”, “Integrative co-occurrence matrices”, “Gray level co-occurrence matrices + color percentiles”, “Gabor features + Chromatic features”, "Gabor Features", “Opponent Color LBP”. These methods are based on texture and disjoint color analysis. Textural features are extracted from images by converting into gray-level and color features are computed with the help of three color component of an image. Textural and color features are concatenated into the same feature vector to improve the classification accuracy.

The main aim of feature selection is to select the maximum number of features to achieve high performance in cancer classification. Feature selection is important when anyone works on gray-level features. In our proposed system, there is no need of feature selection algorithms.

Classification phase of the diagnostic system is the one in charge of making the inferences about the extracted information in the previous phases in order to be able to produce a diagnostic about the input image. In our experiment, we have used four well-established classifiers: Support Vector Classifier (SVC), Nearest Neighborhood (NN), linear classifier, and Nearest Mean Classifier (NMC).

III. PROPOSED ALGORITHM

Proposed framework take image as input, process it and then segmented the processed image using k-mean clustering algorithm. Then color and texture features get extracted from the segments using local binary patterns plus color percentiles and the tested it on different classifier. Finally we checked the accuracy.

Segmentation is a process to partitioning an image into disjoint regions that are homogeneous with respect to a chosen property such as luminance, color, and texture. The aim of segmentation is to change the representation of an image into something that is more meaningful and easier to analyze [8].

A. K-means Clustering Segmentation

K-means clustering is partitioning method. This method groups objects in the way that within-group variance is minimized. If within-group variance is minimized then it gives high featured segmented image. The working of this method is as follows [23]:

- Initialization of any two class centers randomly. These centers represent initial group centroids.
- Calculate the value of histogram bin value distance between each image pixel and class centroids; assign each image pixel to its nearest class centroid.
- Recalculate the new positions of centroids by calculating the mean histogram bin value of the same group.
- If the value of centroids changes then repeat steps b and c.
B. Feature Extraction

It is important to identify the most effective features to extract from skin cancer lesions and to find the most effective pattern-classification criteria and algorithms for differentiating those lesions.

C. Local Binary Patterns + Color Percentiles

The LBP feature vector is created in the following way:

- Divide the window into cells (8x8 pixels for each cell).
- For each pixel in a cell, compare the pixel to each of its 8 neighbors. Follow the pixels along a circle (clockwise or counter-clockwise).
- Where the center pixel’s value is greater than the neighbor’s value, write "1". Otherwise, write "0". This gives an 8-digit binary number (which is usually converted to decimal for convenience).
- Compute the histogram, over the cell, of the frequency of each "number" occurring.
- Normalize the histogram.
- Concatenate histograms of all cells. The output of this process is feature vector for the window.

The LBP; R operator produces 2P different output values where P is the number of neighboring pixels and R is the radius. The combination of LBP+CP presented by Niskanen et al. [7]. In this method, we calculate local binary patterns in a gray scale image. For better results we reduce the calculated features into rotationally invariant features. Rotational invariance is necessary because when image is rotated, the gray values also rotated in a circular form where origin is same. Then calculation of feature vector is changes with the rotation.

D. Identify the Headings

Classification phase of the diagnostic system is the one in charge of making the inferences about the extracted information in the previous phases in order to be able to produce a diagnostic about the input image. There are two different approaches for the classification of dermoscopic images: the first considers only a dichotomous distinction between the two classes (melanoma and non-melanoma) and assigns class labels 0 or 1 to data item. The second attempts to model P (xjy); this yields not only a class label for a data item, but also a probability of class membership. The most prominent representatives of the first approach are support vector machines. Logistic regression, artificial neural networks, k-nearest neighbors, and decision trees are all members of the second approach, although they vary considerably in building an approximation to P (xjy) from data.

In this research, we considered four well-established classifiers, namely: Nearest Neighborhood (NN), Nearest Mean Classifier (NMC), linear classifier, and Support Vector Classifier (SVC).

- An unknown pattern is assigned the label of the nearest training pattern (NN) or that of the nearest centroid (NMC). Absence of parameter tuning makes these classifiers advantageous and easy to implement.
- Linear classifiers classify features by making a classification decision which is based on linear combination of the feature values. The linear classifier is originally developed for binary classification. It requires a predefined linear function (hyper-plane) that best separates the required classes in the feature space. If the two classes are linearly separable then perfect separation between classes exists.
- Support Vector Machine is highly effective classifier, and is currently has a great im-portance in pattern recognition and artificial intelligence. The tuning of a SVM is very important and need very careful analysis. In our experiment, we are using RBF kernel function in SVM classifier then we have to tune two parameters: C and gamma (the radius of RBF).

IV. IMPLIMENTATION AND PERFORMANCE EVALUATION

In our experiment we use database of skin cancer images (Melanoma and Non-melanoma) have been collected from University of Wateloo. University collected these images from internet source. Collection of 150 images is used in this experiment. 75 images are of melanoma and non-melanoma types each respectively. First we follow the pre-processing step in which we resize the image in 128*128 size. We observe that
this size of images optimizes the performance. Matlab is easy to implement and computationally cheap than other simulation software. Therefore we use Matlab to implement our project.

Based on the results, we compile the overall performance of method on classifiers at different classification rate. On the basis of this Table, we can easily say that LBPnCP, perform better than other color feature extraction techniques.

### TABLE I. ACCURACY OF LBPnCP FEATURE EXTRACTION TECHNIQUE ON DIFFERENT CLASSIFIERS

<table>
<thead>
<tr>
<th>Classifier Rate</th>
<th>N</th>
<th>M</th>
<th>Linear</th>
<th>SVC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(60,15)</td>
<td>65.56</td>
<td>50.00</td>
<td>63.33</td>
<td>62.00</td>
</tr>
<tr>
<td>(55,20)</td>
<td>57.14</td>
<td>61.43</td>
<td>51.43</td>
<td>57.14</td>
</tr>
<tr>
<td>(50,25)</td>
<td>68.00</td>
<td>58.00</td>
<td>62.00</td>
<td>66.00</td>
</tr>
<tr>
<td>(45,30)</td>
<td>80.00</td>
<td>66.67</td>
<td>80.00</td>
<td>83.33</td>
</tr>
<tr>
<td>(40,35)</td>
<td>80.00</td>
<td>70.00</td>
<td>100.00</td>
<td>80.00</td>
</tr>
</tbody>
</table>

Based on the results, we compile the overall performance of method on classifiers at different classification rate. On the basis of this Table, we can easily say that LBPnCP, perform better than other color feature extraction techniques.

### V. CONCLUSION

In this paper we have presented some possible approaches to develop a system for automatic classification of skin cancer.

According to results on color and texture analysis, we are using color and texture descriptors for skin cancer classification which provide good classification accuracy on different classification rate. We have also evaluated the performance of four different classifiers on these features. SVC (Support Vector Classifier) and 1-NN outperform among all others with the same number of features set.

Gabor (inter-channel) extraction method is best among all methods. Proper selection of extraction method and classifier in combination is always important for classification of skin cancer images.

Image contrast and lesion orientation directly affect the performance of the system. In our experiment, we assume that contrast is equal and lesion orientation is same.

### REFERENCES


