CATEGORIZATION OF NON-MELANOMA SKIN LESION DISEASES USING SUPPORT VECTOR MACHINE

Mrs.A.Immagulate
Assisitant Professor,
Department f Computer Science,
A.J.K College Of Arts and Science

Abstract : Non-melanoma is the most common form of skin cancer. Early detection of skin cancer has the possible to reduce mortality and morbidity**.** The main objective of this research work is to focus on Non-Melanoma skin cancers and classify the types of it. There are five types of Non-melanoma skin cancers namely Actinic Keratosis (AK), Basal Cell Carcinoma (BCC), Melanocytic Nevus / mole (MN), Squamous Cell Carcinoma (SCC), and Seborrhoeic Keratosis (SK). The database contains allthe above mentioned five types of skin lesion diseases images and it is used to perform three different stages namely preprocessing the image, feature extraction for each image and classification of skin cancer. In preprocessing stage the input images are resized and adjusted and in feature extraction task 29 different features are extracted for all images in the given database. Finally the classification task is performed using Support Vector Machine (SVM) and Active Support Vector Machine (ASVM). The experimental result shows that Active Support Vector Machine performs well in finding out the Non-Melanoma skin cancer accurately and efficiently.

IndexTerms - Skin cancer, Actinic Keratosis (AK), Basal Cell Carcinoma (BCC), Melanocytic Nevus / mole (MN), Squamous Cell Carcinoma (SCC), Seborrhoeic Keratosis (SK), Support Vector Machine (SVM) and Active Support Vector Machine (ASVM).

I. INTRODUCTION

Skin cancer starts in the cells of the skin. The skin is the body's largest organ. It protects the human being from injury, infection, heat and ultraviolet light from the sun. The skin helps control human being body temperature and gets rid of waste materials through the sweat glands. It also makes vitamin D and stores water and fat. Skin cancer is often categorized into either melanoma or non melanoma. Non melanoma skin cancer occurs in either basal or squamous cells. These cells are located at the base of the outer layer of the skin or cover the internal and external surfaces of the body.

Non melanoma skin cancer is the most common type of cancer. Non melanoma skin cancer is mainly divided into 5 type's namely actinic keratosis, basal cell carcinoma, melanocytic nevus, squamous cell carcinoma, seborrhoeic keratosis. An actinic keratosis, also known as a solar keratosis, is a scaly or crusty growth. It most often appears on the bald scalp, face, ears, lips, backs of the hands and forearms, shoulders, neck or any other areas of the body frequently exposed to the sun. Basal cell carcinoma begins in the basal cells - a type of cell within the skin that produces new skin cells as old ones die off. Basal cell carcinoma often appears as a waxy bump, though it can take other forms. A melanocytic nevus is a type of lesion that contains nevus cells, a type of melanocyte. Some sources equate the term mole with melanocytic nevus. A melanocytic nevus blemish may be a kind of lesion that contains blemish cells. Squamous cell carcinoma is an uncontrolled growth of abnormal cells arising in the squamous cells, which compose most of the skin's upper layers called the epidermis. SCCs may occur on all areas of the body including the mucous membranes and genitals, but are most common in areas frequently exposed to the sun, such as the rim of the ear, lower lip, face, bald scalp, neck, hands, arms and legs. Seborrheic keratosis is a very common harmless, usually pigmented, noncancerous growth on the skin. It usually appears as a pale, black or brown growth on the back, shoulders chest or face, but can appear anywhere on the skin.

Biopsy is the standard diagnosis procedure carried out in diagnosing the skin cancer. Biopsy is a sample of tissue collected from an organ or other part of the body. A biopsy can be done by cutting or scraping a small piece of the tissue or by using a needle and syringe to remove a sample, which is then examined for abnormalities, such as cancer, by a doctor trained to look at tissue samples**.** This procedure of diagnosing the diseases is time consuming and may lead to inaccurate prediction of the disease. Hence, in this research work it is proposed to automate the prediction of skin lesion using machine learning approach. Based on the study of various literatures available on non melanoma skin lesion categorization, a brief report is presented in this section about the developments in categorization of non melanoma in the last several years.

Lucia Ballerina and Robert B. Fisher et al., [2] have used *k*-nearest-neighbor classification. Watershed segmentation was the main pre-processing steps. Color and texture features were extracted from the skin lesion images. The image database comprised of 960 lesions that belonged to 5 classes. *K-nearest-neighbor* classification technique gave 74% accuracy. But difficulty of this method was segmentation of skin lesion becomes difficult in the presence of shadows and bright areas caused by illumination variation.

M. Emre Celebi and Hassan and A. Kingravi et al., [3] proposed a methodological approach to the classification of pigmented skin lesions in dermoscopy images. First, automatic border detection was performed to separate the lesion from the background skin. Shape features were then extracted from the border. The SVM classification was used for the classification of the images. Experiments on a set of 564 images yielded a specificity of 92.34% and a sensitivity of 93.33%. But the disadvantages of this method were an obtained boundary might merge with another boundary due to image noise and produce a false lesion boundary. If parameter was too small, some parts of the optimal boundary might be missed, while if it was too large, detection of the optimal boundary might become difficult.

José Fernández Alcón, Calina Ciuhu, Warner ten Kate, Adrienne Heinrich et al., [4] proposed Automatic Imaging System With Decision Support for Inspection of Pigmented Skin Lesions and Melanoma Diagnosis. The system included a dedicated image processing system for feature extraction and classification, and patient-related data decision support machinery for calculating a personal risk factor. A robust segmentation algorithm had been developed. Texture feature was extracted from the image. And this method produced 81% of accuracy.

In this research support vector machine and active support vector machine have been adopted for constructing the non melanoma skin lesion prediction model. The SVM based trained model will facilitate accurate prediction of the skin lesion images.

II. PROPOED WORK

The skin lesion images related to five types of diseases such as actinic kertosis, basal cell carcinoma, melanocytic nevus, squmous cell carcinoma, sebhorreic kertosis have been collected from two different websites namely Dermnet.com and DERMOFIT.com. The database contains images of five types of diseases each having 100 samples. The images are preprocessed using image adjustment and image resize. The color and texture features are extracted from the preprocessed skin lesion images and training dataset is developed. The training dataset is trained using multiclass support vector machine and its variants such as proximal support vector machine and active support vector machine. In modeling skin lesion prediction the essential tasks such as data acquisition, preprocessing, feature extraction, model building and testing are carried out.

The architecture of proposed non melanoma skin lesion categorization is shown in Fig.3.1.
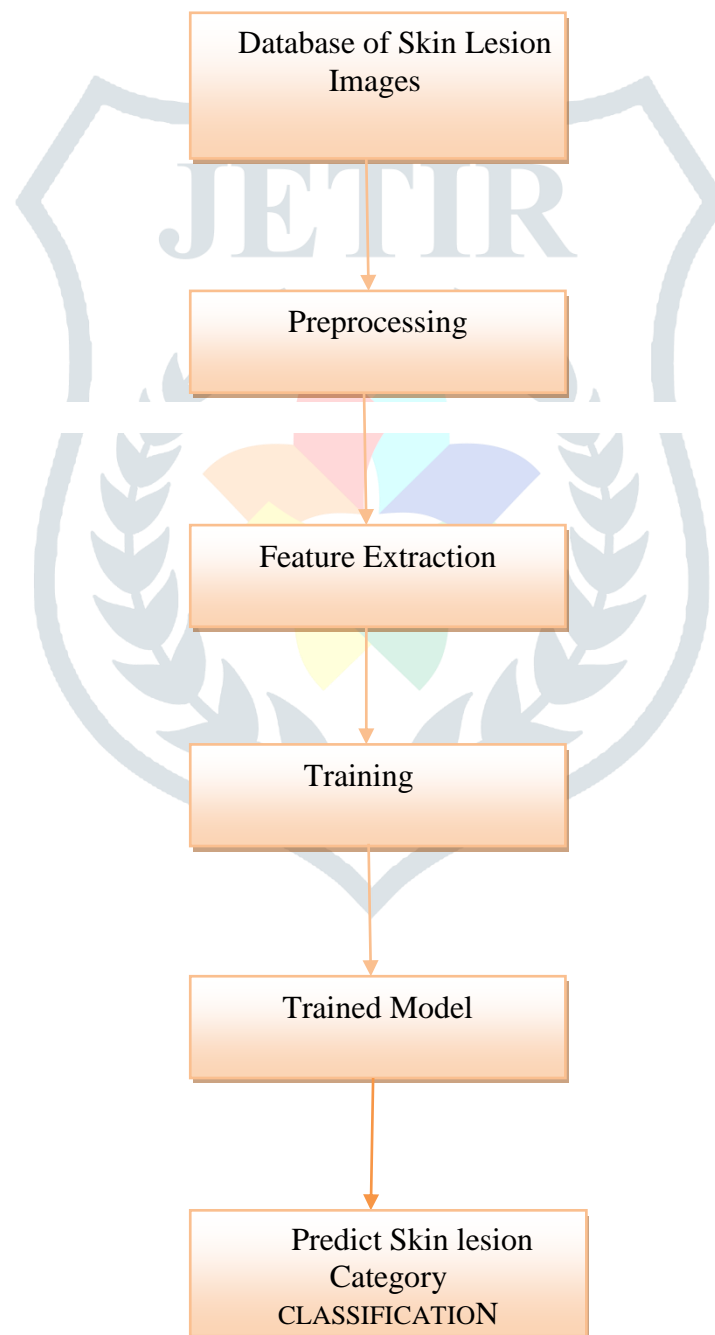

Fig. 1 Proposed non melanoma skin   lesion categorization

2 .1 DATA ACQUISITION

The data preparation is an important task in any machine learning activity. The images related to five types of non melanoma skin lesions have been collected from Dermnet.com and DERMOFIT.com. Five sets of images namely actinic kurtosis, basal cell carcinoma, sqmous cell carcinoma, melanocytic nevus, sebhorric kurtosis, each consisting of 100 images are taken into consideration. The sample images are shown in Fig.2.2.

Acitinic Kertosis　　　Basal Cell Carcinoma

Melanocytic Nevus　　　Sebhorreic Kertosis

Squmous Cell Carcinoma
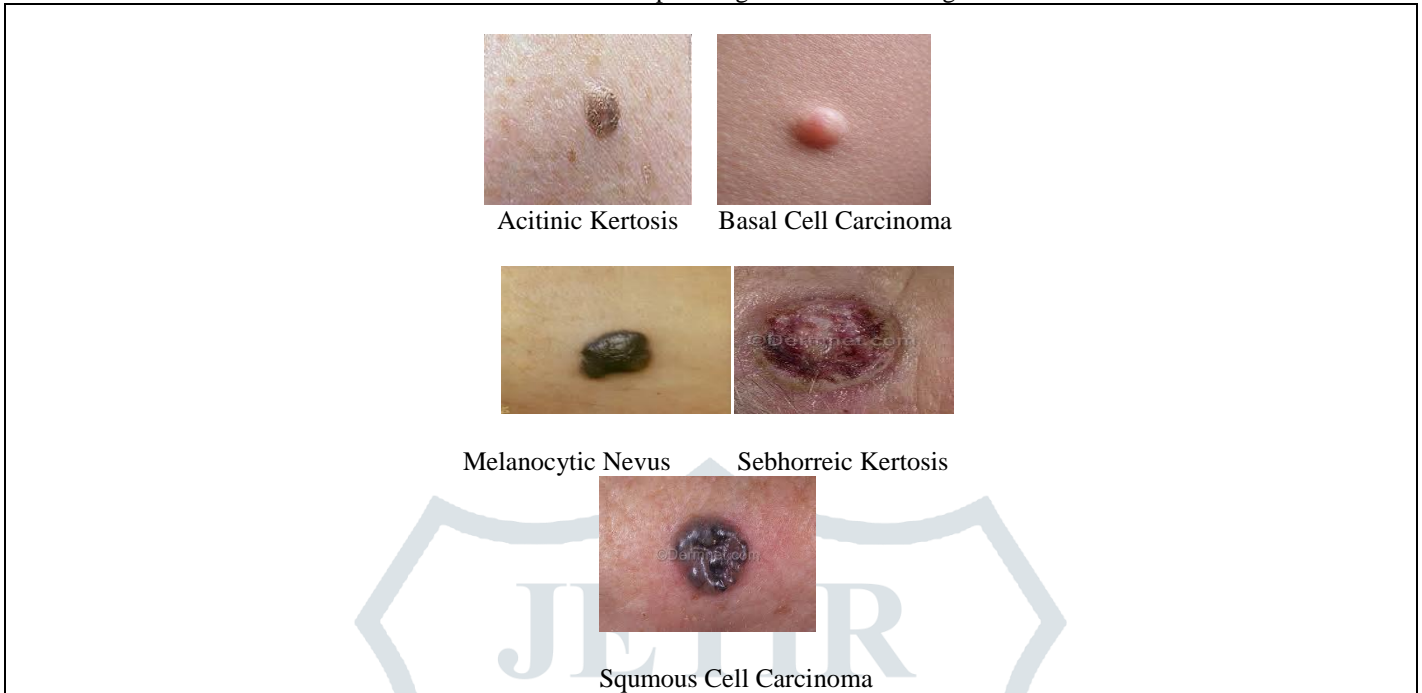
Fig.2.2. Sample images of non melanoma skin lesion

2.2 PREPROCESSING

The aim of preprocessing is to improve the image data that suppresses unwanted distortions for further processing. Preprocessing is the process of removal of noise and transformation of the image to produce a well defined pattern for texture analysis. In this work, two steps of pre processing are performed. The first step is image resize and the second is image adjustment.

2.2.1 Image resize

As the size of the images vary for different images, it is essential to make them uniform. The images are resized to 256*256 dimensions using interpolation method. Interpolation is the process used to estimate an image value at a location in between image pixels. When imresize function of matlab is used, the image enlarges and the, output image contains more pixels than the original image.

2.2.2 Image adjustment

This preprocessing step is mainly used to adjust the image intensity values or color map values. As color features are based on the color pixel rate, the image adjustment is required to change the color pixel values into correct and accurate format. Another function carried out in this step is to adjust the intensity of the color images. It brings down the high intensity and increases the low intensity of the images to convert them into a medium intensity images.

III. FEATURE EXTRACTION

Feature extraction plays a vital role in data mining. It can be used to improve the classification effectiveness and computational efficiency. Drawing out specific features from the preprocessed images is called feature extraction. Feature extraction is carried out with all the preprocessed skin lesion images. The features describe the distinguishing pattern of the skin lesion images. Two types of features such as color and texture features are extracted.

3..1 Color features

Color features play a decisive role in the classification of skin lesions. A color image is a combination of some basic colors. Each individual pixel of a color image is broken down into red, green and blue values.

The RGB color features such as meanR, meanG, meanB, normalizedR, normalizedG, normalizedB, are extracted from skin lesion images. First a color image is split into three separate RGB color planes using the following functions (2).

Image_red = Image_rgb(:,1);
Image_green=Image_rgb(:,2);
Image_blue = Image_rgb(:,3);

Then the red, green and blue pixel values are calculated. Finally the mean values of the color planes R, G and B are calculated using the following functions.
mean_r=mean2(Red);
mean_g=mean2(Green);
mean_b=mean2(Blue);

Using minmax normalization, the meanR, meanG, meanB features are normalized and normalizedR, normalizedG, normalizedB features are computed. The following min max normalization is used for normalization.

$$v' = \frac{v - minA}{maxA - minA}(new\_maxA - new\_minA) + new\_minA$$

b. Texture Features

The texture features are extracted using Grey Level Co-occurrence Matrix (GLCM). A GLCM is a matrix where number of rows and columns is equal to number grey levels G in a image. It is defined over an image to be the distribution of co-occurring values in the given offset. It is a way of extracting second order statistical features. It is used to measure the spatial relationships between pixels. This method is based on the belief that texture information is contained in such relationships. Few of the common statistics applied to co-occurrence probabilities are given the following table with related formulas;

Table: Some of the texture features

Feature	Formula
Energy	$$energy\ (ene) = \sum_i \sum_j {g_{ij}}^2$$
Entropy	$$entropy\ (ent) = -\sum_i \sum_j g_{ij} \log_2 g_{ij}$$
Contrast	$$contrast\ (con) = \sum_i \sum_j (i-j)^2 g_{ij}$$
Variance	$$variance\ (var) = \sum_i \sum_j (i-\mu)^2 g_{ij}\ where\ \mu\ is\ the\ mean\ of\ g_{ij}$$
Homogeneity	$$homogeneity\ (hom) = \sum_i \sum_j \frac{1}{1+(i-j)^2} g_{ij}$$
Correlation	$$correlation\ (cor) = \frac{\sum_j \sum_j (ij)g_{ij} - \mu_x \mu_y}{\sigma_x \sigma_y} g_{ij}$$
Autocorrelation	$$p(x,y) = \frac{\frac{1}{(L_x-|x|)(L_y-|y|)} \iint_{-\infty}^{\infty} I(u,v)I(u+x,u+v)du\,dv}{\frac{1}{L_x L_y} \iint_{-\infty}^{\infty} I^2(u,v)du\,dv}$$
Sum Average	$$sum\ average\ (sa) = \sum_{i=2}^{2N_g} i g_{x+y}(i)$$
Sum Entropy	$$sum\ entropy\ (se) = -\sum_{i=2}^{2N_g} i g_{x+y}(i) \log\{g_{x+y}(i)\}$$
Sum Variance	$$sum\ variance\ (sv) = \sum_{i=2}^{2N_g} (i-sa)^2 g_{x+y}(i)$$
Difference variance	$$difference\ variance = variance\ of\ g_{x-y}$$

Difference Entropy	$difference\ entropy\ (se)$ $$= -\sum_{i=0}^{N_g-1} g_{x-y}(i)\log\{g_{x-y}(i)\}$$

IV. CLASSIFICATION

4.1. Support vector machine

Support vector machines [7] (SVMs) are a set of related supervised learning methods used for classification and regression. A support vector machine constructs a hyper plane or set of hyper planes in a high-dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class called functional margin, since in general the larger the margin the lower the generalization error of the classifier.

Let's introduce the notation used to define formally a hyperplane:

$$f(x) = \beta_0 + \beta^T x$$

where β is known as the weight vector and β_o as the bias. The optimal hyperplane [8] can be represented in an infinite number of different ways by scaling of β and β_o. As a matter of convention, among all the possible representations of the hyperplane, the one chosen is

$$|\beta_0 + \beta^T x| = 1$$

Where x symbolizes the training examples closest to the hyperplane. In general, the training examples that are closest to the hyperplane are calledsupport vectors. This representation is known as the canonical hyperplane. Now, the result of geometry that gives the distance between a point x and a hyperplane (β,β_o):

$$Distance = \frac{|\beta_0 + \beta^T x|}{\|\beta\|}$$

In particular, for the canonical hyperplane, the numerator is equal to one and the distance to the support vectors is

$$Distance_{support\ vectors} = \frac{|\beta_0 + \beta^T x|}{\|\beta\|} = \frac{1}{\|\beta\|}$$

Recall that the margin introduced in the previous section, here denoted as , is twice the distance to the closest examples:

$$M = \frac{1}{\|\beta\|}$$

Finally, the problem of maximizing M is equivalent to the problem of minimizing a function L (β) subject to some constraints. The constraints model the requirement for the hyperplane to classify correctly all the training examples x_i. Formally,

$$min_{\beta,\beta o}L(\beta) = \frac{1}{2}\ \|\beta\|^2\ subject\ to\ y_i(\beta^T x_i + \beta_0) \geq 1\forall_i$$

where y_i represents each of the labels of the training examples.

4.2 . Active Support Vector Machine (ASVM)

The algorithm [9] consists of determining a partition of the dual variable u into non-basic and basic variables. The non-basic variables are those which are set to zero. The values of the basic variables are determined by finding the gradient of the objective function of SVM with respect to these variables, setting this gradient equal to zero, and solving the resulting linear equations for the basic variables. If any basic variable takes on a negative value after solving the linear equations, it is set to zero and becomes non-basic. This is the essence of the algorithm. In order to make the algorithm converge and terminate, a few additional safeguards need to be put in place in order to allow us to invoke the More-Toraldo finite termination result.

The other key feature of the algorithm is a computation a l one and makes use of the SMW formula. This feature allows us to invert an (n + 1) x (n + 1) matrix at each step instead of a much bigger matrix of order m x m.

H = D[A - e], Q = I /v + HH'.

With these definitions the dual problem becomes

$$min_{0\leq u\in R^m}f(u):= \frac{1}{2}\ u'Qu - eu$$

It will be understood that within the ASVM Algorithm, Q - 1 will always be evaluated using the SMW formula and hence only an (n+l) x (n+l) matrix is inverted.

this new notion of error.

IV. EXPERIMRNTS AND RESULTS

The classification algorithms are implemented in Matlab. Three experiments have been carried out by implementing support vector machine and its variants. The training dataset is developed using 500 images related to five types of skin lesion diseases. Image resize and image adjustment are the two main preprocessing tasks carried out. The color and texture features are extracted from the preprocessed skin lesion images to form feature vectors. For each feature vector, class labels 1 to 5 are assigned. The comparative results indicate that PSVM based classification model yields a better performance when compared to other models. The comparative results of SVM, ASVM, and PSVM in terms of accuracy and learning time is shown in the following table;

Table: Evaluation measures

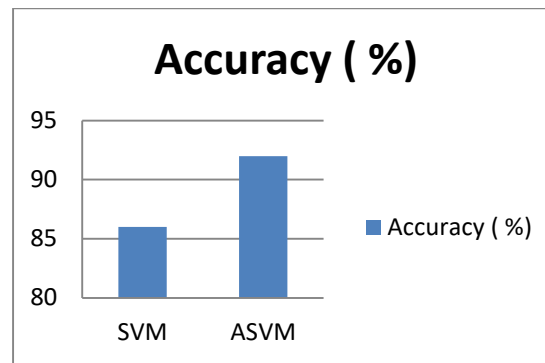| Classifier	Accuracy (%)	Learning Time
SVM	86	85
ASVM	92	60

Accuracy (%)

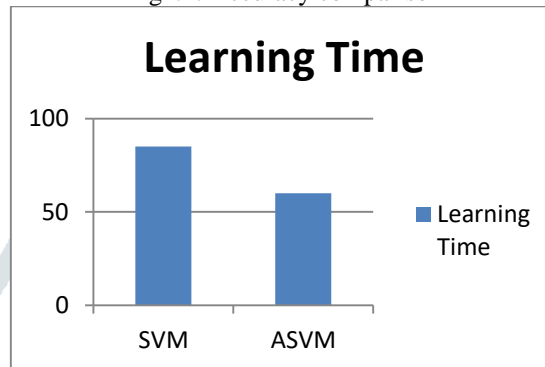Fig4.1: Accuracy comparison

Learning Time

Fig4.2: Learning time comparison

The above figures (fig4.1 & 4.2) illustrate the comparison of accuracy and learning time of each classifiers. As far as the skin lesion prediction model is concerned, predictive accuracy plays a major role than learning time. Once the model is developed, it can be integrated with medical diagnosing system. Therefore it is not necessary to pay much attention to learning time.

V. CONCLUSION

In this research, an automated medical decision support system for non-melanoma skin cancer developed with five different classes. First the preprocessing is applied to get resized and adjusted images and the GLCM is applied for extracting various types of features from each images. Finally classification algorithms such as support vector machine and its variants are applied to classify the skin lesion cancer. The comparative results indicate that active support vector machine based non melanoma skin lesion categorization model exhibit   better performance when compared to SVM. The results of the deployed techniques were promising as, 92% of classification accuracy with the minimum learning time.

REFERENCES

[1]. G. R. Day & R. H. Barbour. "Automated melanoma diagnosis: where are we at?" Skin Research and Technology **6**, pp. 1–

5,2000.

[2]. Lucia Ballerina, Robert B. Fisher Ben Aldridgey, Jonathan Reesy "Non-melanoma skin lesion classification using colour image data in a Hierarchical k-nn classifier"

[3]. M. E. Celebi, W. V. Stoecker, and R. H. Moss, "Advances in skin cancer image analysis," Computerized Medical Imaging and Graphics, vol. 35, No. 2, Pp. 83 – 84, 2011.

[4] M. Emre Celebi and Hassan and A. Kingravi"A methodological approach to the classification of dermoscopy images"Comput Med Imaging Graph. 2007 September; 31(6): 362–373.

[5]José Fernández Alcón, Calina Ciuhu, Warner ten Kate, Adrienne Heinrich, Natallia Uzunbajakava, Gertruud Krekels, Denny Siem, and Gerard de Haan-2009"Automatic Imaging System With Decision Support for Inspection of Pigmented Skin Lesions and Melanoma Diagnosis"

[6].I. Maglogiannis And C. N. Doukas, "Overview Of Advanced Computer Vision Systems For Skin Lesions Characterization," Ieee Transactions On Information Technology In Biomedicine, Vol. 13, No.,Pp.721–733,2009.

[7]. 14. Ruben Nicolas,1 Albert Fornells," DERMA: A Melanoma Diagnosis Platform Based onv Collaborative Multilabel Analog Reasoning".

[8]. 18. Cortes, Corinna, Vapnik and Vladimir N, "Support Vector Networks, Machine Learning".

[9].Xin Li and Yuhong Guo "Active Learning with Multi-Label SVM Classification" Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence.

[10]. Andreas Vlachos," Active Learning with Support Vector Machines

[11]." GLENN M. FUNG," Multicategory Proximal Support Vector Machine Classifiers" 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands.