Dermatological Disease Detection Using Image Processing and Machine Learning

Vinayshekhar Bannihatti Kumar
Computer Science Department
PES Institute of Technology
vinayshekhar000@gmail.com

Sujay S Kumar
Computer Science Department
PES Institute of Technology
sujay.skumar141295@gmail.com

Varun Saboo
Computer Science Department
PES Institute of Technology
v18saboo@gmail.com

***Abstract*— Dermatological diseases are the most prevalent diseases worldwide. Despite being common, its diagnosis is extremely difficult and requires extensive experience in the domain. In this research paper, we provide an approach to detect various kinds of these diseases. We use a dual stage approach which effectively combines Computer Vision and Machine Learning on clinically evaluated histopathological attributes to accurately identify the disease. In the first stage, the image of the skin disease is subject to various kinds of pre-processing techniques followed by feature extraction. The second stage involves the use of Machine learning algorithms to identify diseases based on the histopathological attributes observed on analysing of the skin. Upon training and testing for the six diseases, the system produced an accuracy of up to 95 percent.**

***Keywords*— Dermatology, Image Processing, Computer Vision, Machine Learning, Data Mining, Computational Intelligence, Automated Disease Diagnosis**

I. INTRODUCTION

Dermatology is one of the most unpredictable and difficult terrains to diagnose due its complexity. In most developing countries, it is expensive for a large number of people to consult a dermatologist. The ubiquitous use of smart phones in a developing country has opened up new avenues for inexpensive diagnosis of diseases. We can use the camera technology present in every smartphone and exploit the image processing capabilities of the device for diagnosis. We have developed an application that utilizes a two staged approach in order to tackle the problem. The first stage involves Image Processing for identification and the second stage involves Machine Learning for a near fool proof solution. Difficulty for the differential diagnosis is that a disease may show the features of one disease in the initial stage and may have the characteristic features of another in the following stages. Usually a biopsy is necessary for the diagnosis but these diseases share many histopathological features as well. This issue is solved by using machine learning models on the clinically evaluated features which are determined by an analysis of the skin samples under the microscope.

Owing to the subjective nature of diagnosis, medical students find it difficult to verify their diagnosis. This system acts as an effective learning tool, aiding verification of their results as they have access to clinical data. The training data set was obtained from the machine learning data repository of University of California, Irvine [18]. We have achieved higher accuracies using an ensemble of Computer Vision and Machine Learning algorithms. The system is capable of detecting six of the most commonly occurring diseases, namely – Psoriasis, Seborrheic Dermatitis, Lichen Planus, Pityriasis Rosea, Chronic Dermatitis, and Pityriasis Rubra Pilaris.

II. PAST WORK

The paper proposed by Muhammad Zubair Asghar et al [16] put forward a rule based web supported expert system to detect certain skin diseases using forward chaining with depth first searching. However, using a rule based system in order to detect the type of dermatological condition is not practical due to the various manifestations of a single skin disease. A self-learning model developed by us would be a better performer in this regard as the problem we are trying to address is probabilistic in nature and hence we need a system which learns the underlying pattern present in the skin disease which can be inferred by the image and the histopathological inputs.

Rahat Yasir et al [2] have proposed a self-learning system which is capable of detecting a skin disease by using image processing and artificial neural networks. Although this is an effective way of detecting a skin disease using an image, we still feel there is inadequacy in the number of features extracted from the image itself. For a successful diagnosis of a skin disease, there is a need to involve other histopathological attributes.

Although the system proposed by A.A.L.C. Amarathunga et al [1] proceeds in the direction of including a data mining unit to the system of skin detection, it is lacking in the choice of attributes considered for detection. Neither the data source nor the attributes used for learning/testing has been mentioned.

Kabari et al [13] have used an artificial neural network system which diagnoses skin diseases and have been able to achieve accuracy of 90%.

M. SHAMSUL et al [14] have also proposed an image processing system with pre-processing algorithms and a feed forward neural network similar to [2].

Shuzlina et al [15] have implemented a back propagation neural network which resulted in an accuracy of 91.2%.

Florence et al [20] classifies the image as a bacterial or viral skin infection using image processing techniques.
Damilola et al [21] has modelled a system that collates pigmented skin lesions image results, analysis, corresponding observation and conclusions by medical experts using prototyping methodology. It uses computational intelligence technique to analyze, process and classify the image library data based on texture.

III. ARCHITECTURE AND METHODOLOGY

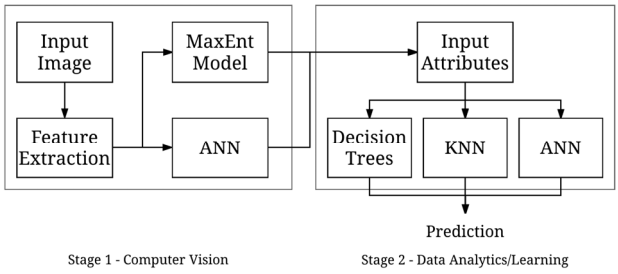

Fig.1. Block Diagram

A. Stage 1 - Computer Vision

The system uses Computer Vision as the first stage in identification of the type of skin disease based on the numerous features extracted from the image using various image processing techniques. The computer vision stage itself consists of two phases. In the first phase, we pre-process the image taken through the camera of the smart phone in order to extract the necessary features. The second phase involves using the features extracted in order to identify the disease using various algorithms like Maximum Entropy Model and Artificial Neural Networks. The features that were extracted includes the color code of the inflicted area, size of infliction, contrast of the infliction with respect to the surrounding the healthy skin, shape through edge detection of the infected area.

B. Stage 2 – Machine Learning

The system uses various Machine Learning techniques in the second stage in order to refine the classification of the image. The training dataset for this stage was sourced from the UCI dataset [18]. The second stage of prediction is made available to the medical professionals who have access to various histopathological attributes like exocytosis, hyperkeratosis, acanthosis, parakeratosis and other attributes as mentioned in [18]. The system takes these attributes as input from the user and gives a better classification of the disease. We train the system using the dataset obtained from the UCI repository [18] and test it using Decision Trees, Neural Networks and kNN (kth Nearest Neighbour) Model.

IV. DATA SOURCE

A. Computer Vision

The images for training were obtained from MS Ramaiah Medical College, Bangalore and Mysore Medical College, Mysore, with the help of Dr. Supriya Ramesh and Surabhi Sainath respectively. The tagged sets of images were taken from the Department of Dermatology.

B. Machine Learning

The dataset used for training was obtained from the learning data repository of University of California, Irvine [18].

V. COMPUTER VISION

A. Image Pre-processing

The colour images obtained from the phone camera is pre-processed before using it for feature extraction. In order to improve the accuracy of feature extraction, eight different pre-processing algorithms were used. The algorithms used were converting to grey scale image, sharpening filter, median filter, smooth filter, binary mask, RGB extraction, histogram and sobel operator. The RGB values of the images is extracted before converting it into a grayscale image. Sharpening filter is applied to the grayscale image in order to sharpen the details of the infected region. Median filter is used after sharpening filter in order to remove the noise from the image. The next algorithm used is the smoothing filter which replaces each pixel with the mean values of its neighbours, including itself. Binary image was generated from the mean filtered image and distribution of colour of binary image was showed by histogram. YCbCr was used to extract average colour code of the infected area from the binary image. Sobel operator was applied to binary image to detect edge of the infected area.

B. Feature Extraction

The first feature extracted from the color image is the color histogram of the inflicted region of the skin. We converted RGB color space of the image into HSV color space, as outlined in [8]. HSV color space was preferred over the RGB color space as it separates color components (HS) from the luminance component (V) and is less sensitive to changes in illumination of the image.
The Sobel operator was used to detect edges which helped us in detecting the shape of the infliction. This was done by using the various image segmentation algorithms including Otsu's method, Gradient Vector Flow and color based image segmentation as outlined in the paper [12].
The number of components of the skin affliction was extracted from the image using the Euler value. A threshold limit was imposed on the Euler value heuristically, exceeding which was an indicator of presence of a large number of inflictions. This is an important distinguishing feature characteristic for diseases such as pityriasis rosea and dermatitis, which has been neglected in the [2].
Another important feature that we extracted was the presence or absence of bumps around the hair follicles which is the distinguishing feature of the disease pityriasis rubra pilaris. This feature was extracted by first detecting the presence of hair in the image as described by the DullRazor algorithm

[22]. If hair is present, an oval shaped patch is detected and checked in order to determine the presence of bumps.
The other features including detecting if the affected region was face, scalp or torso and knees were extracted by detecting distinguishing features like nose, eyes, mouth, hairline, ears or navel in order to distinguish between involvement of face and torso/knees.
The next feature that we extracted was the presence/absence of patches in the nail beds. For this, we used the fingernail detection method from hand images as explained in [6], using not only distribution density, but also colour continuity, for improved accuracy.
The next feature extracted was the presence/absence of pustules in the afflicted skin image. This was extracted by considering both the number of components extracted before and the colour variation using colour homogeneity [3].
The presence/absence of scaling in the affected region was also extracted. This was done by using the texture analysis as outlined in [9].

C. Probabilistic Distribution of Diseases From Features

The features were taken into consideration along with the diseases and given a probability distribution based on MaxEnt model.

Feature Functions Table

	Facial Presence	TKE	Nails	Color (Purple)	Plaque (Scaling)	Pustular	Hair Loss	CG5	Bumps	Oval Shaped
I	1	1	1	0	1	1	0	0	0	0
II	1	0	0	0	1	0	1	1	0	0
III	1	0	0	1	0	0	0	0	0	0
IV	0	1	0	0	0	0	0	1	0	1
V	1	1	0	0	1	0	0	0	0	0
VI	0	1	0	0	1	0	0	0	1	0

Table 1

TKE - Torso, Knee or Elbow Presence
CG5 – Number of Components > 5
(*Roman numerals indicate Disease)*

Identification of Disease from the features

1) Psoriasis: It can be identified based on the occurrence of plaques on the face, torso, knees and elbows. The special feature of this disease is that it occurs near the nail beds as well.
2) Seborrheic Dermatitis: This condition is mainly seen on the head. Hence, the two values corresponding to these feature functions are likely to be set to 1.
3) Lichen Planus: Being present on the face, it is harder to differentiate from other diseases, but it exhibits purple coloured patches which turns on its respective feature function.
4) Pityriasis Rosea: This condition affects mainly the knees and elbows. The number of components becomes a distinguishing feature as the number of rashes is high.
5) Chronic Dermatitis (CD): Although this condition affects both face and torso, it can be distinguished with the presence of distinct flaking.
6) Pityriasis Rubra Pilaris: This condition is known to affect only the torso and lower body. The distinguishing feature is the presence of bumps around hair follicles.

D. Machine Learning Models Used for Prediction

Previous attempts at classification of diseases did not lead us to great accuracy and required a lot of training time with the use of Artificial Neural Networks as proposed in [2]**.** Since the need was to get high computing which could be used by mobile devices, there was a need to decrease the computations required for training and testing. Another problem with [2] is that the number of features extracted from the image itself is very less, making it more human dependent and less automated. We, on the other hand, extracted the mentioned features and passed it to a Maximum Entropy model which gave us better results because of the number of feature functions and its applicability to this problem, which not just being mutually exclusive, but were also approved by a certified medical professional.

1) Maximum Entropy Model:
Consider the principle of maximum entropy as stated by Jaynes, in making inferences on the basis of partial information we must use that probability distribution which maximises entropy subject to whatever is known.

$$\mathrm{H(P)} = -\sum_x (p(x) log(p(x)) \quad (1)$$
$$where\ \mathrm{x=(a,b)a}\epsilon A, b\epsilon B \text{ and } E = A \times B$$

$$v.f(x,y) = \sum_{k=1}^{m} v_k f_k$$

Here v is the model parameter.
The conditional Log likelihood is used to train the model parameters as described in**.** Theta as described by Kevin Gimple et al [4] is the model parameter (vector).

$$\min_\theta \sum_{i=1}^{n} -\theta^T f(x^{(i)}, y^{(i)}) + log \sum_y exp(\theta^T f(x,y))$$

The output is computed by using the below equation and picking the disease with highest probability.

$$p(y = Disease|x = TestExample) = \frac{exp(vf(x,y))}{\sum_{y'} exp(vf(x,y'))}$$

2) Artificial Neural Network:
A feed forward ANN with back propagation was also tried along with the Maximum Entropy model for the same features which were extracted. The ANN consisted of one input layer, two hidden layers and an output layer.
- Input Layer: The features extracted.
- Hidden Layer 1: Composed of sigmoid neurons

$$\sigma(\theta^T X) = \frac{1}{1+e^{\theta^T X}}$$

where θ is the model parameter,X is the input example and let A be the output of this layer

- Hidden Layer 2: Composed of tanh function

$$\tanh(\theta 1^T X) = \frac{e^{\theta 1^T X} - e^{-\theta 1^T X}}{e^{\theta 1^T X} + e^{-\theta 1^T X}}$$

  Here X is A, as the output of the sigmoid layer is fed as input to the tanh layer. Let the output of this layer be a vector B.
- Output Layer : Softmax Layer

$$p(y|X) = \frac{exp(\theta^T B)}{\sum_{i=1}^{n} exp(\theta^T B)}$$

  Here, B is the output obtained after passing through the tanh layer.

[2] has implemented a yes/no output layer which is not the optimal solution in our case for two reasons.

- We need to classify the image into one of the six diseases and this is not possible with a yes/no output layer.
- We cannot use a model which assumes pre-defined prior probabilities.

With the implementation of softmax layer as output, we were able to get a probability distribution across all six diseases. The argmax of this distribution gave us the name of the disease.

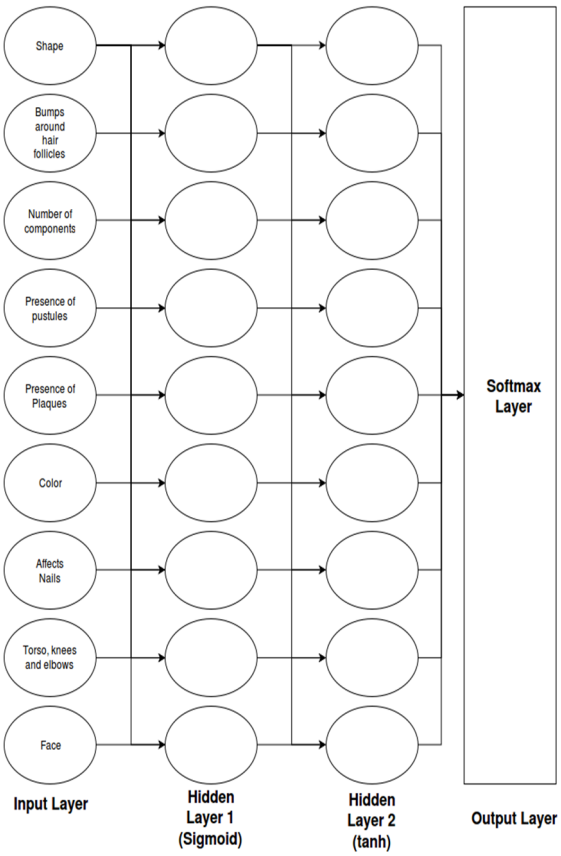

Fig. 2. Artificial Neural Network used for prediction

E. Computer Vision Results

Fig. 3. Test image for Psoriasis

Fig. 4. Test Image for CD

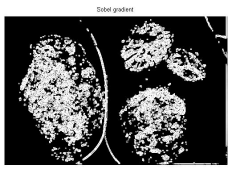

Fig. 5. Edge detected

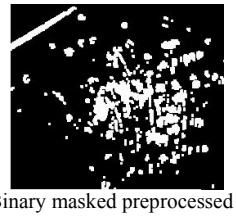

Fig 6. Binary masked preprocessed image

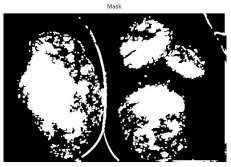

Fig 7 .Binary Image

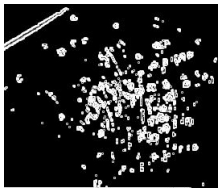

Fig 8 .Sobel operator for CD

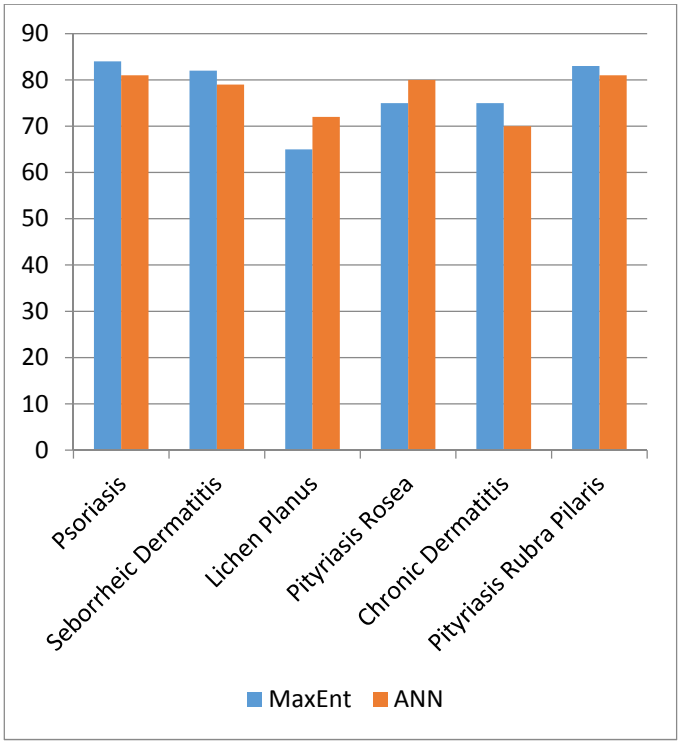

Fig. 9. Computer Vision Results

There was a difference in accuracy between the two models which were used for testing, as can be seen in [Figure 6]. The results were as predicted. [Figure 3] is a sample input image depicting Psoriasis and [Figure 4] and [Figure 5] are the intermediate images produced after image pre-processing. The hand coded rules as suggested by a dermatologist for feature functions helped us get a better accuracy with the use of the

maximum Entropy model. We took a hit on the accuracy for detecting Lichen Planus due to its varied manifestations because of which the dot product of the parameter vector and the feature function itself was minimal ,that is v.f(x,y)=0.

VI. MACHINE LEARNING

A. Motivation for Machine Learning

There are mainly two types of users for this app, naïve users who do not have access to the extremely technical details of histopathological attributes and those users who are aware of this data because they have the access to the medical reports generated by the laboratories. We exploit this fact in order to improve the efficiency of the system. These medically approved attributes are better performers when compared to the features extracted from the image alone.

B. Data Source

The dataset obtained from [18] is used to train the model. There were various models which were tested in order to learn the hypothesis which best suits this training data.

C. Machine Learning Models Used for Prediction

1) kth Nearest Neighbor (kNN)

A non-parametric algorithm used for classification. The various test data are considered to be points on a 34 dimensional space, because there are 34 attributes. When the new test data comes in, the k closest points to this are put in a poll and the result gives the label of the disease.

2) Decision Trees (DT)

The algorithm which was used to train the decision tree was the ID3 algorithm [10][17]. There are two advantages of using the decision tree. The decision tree is the best suited algorithm for training a system with missing values. The ID3 algorithm works on the principle of choosing the attribute with lowest entropy (1) and highest information gain [5].

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

The attributes which have missing data will have very low information gain. For this data set the attributes with missing data were found not found in the tree as the information gain was too low on these attributes.

3) Artificial Neural Networks (ANN)

Any non-linear system can be built with just one hidden layer but must have 2^n units in hidden layer; this is known as universal approximator. [19] This is not practical as described by George Cybenko. A feed forward ANN with back propagation was tried for the same features which were extracted. The ANN consisted of one input layer, two hidden layers and an output layer.

A deep learning technique called auto encoder [11] which trains the model using an identity function was used to initialize the parameters of the model, the weight matrices W.

- Input layer: Input from the user (34 neurons)
- Hidden Layer 1: Composed of sigmoid function (16 neurons)
- Hidden Layer 2: Composed of tanh function (8 neurons).
- Output Layer: Softmax as described by (1).

One issue with using ANN was the possibility of under fitting. Due to the presence of missing values in the training dataset [18] the artificial neural network does not suit well for this problem as there will be too much under fitting.

D. Results

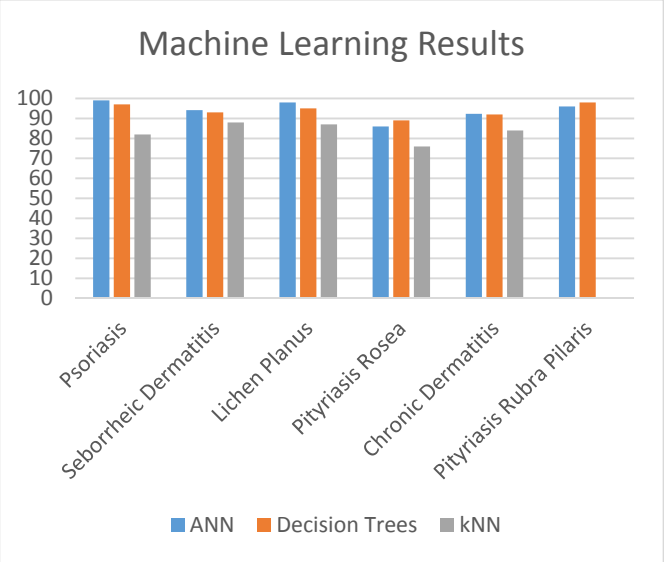

Fig. 10. Machine Learning Results

To obtain these results, 67% of the data obtained from [18] is used for training and 33% is used for testing the models. The results as shown in [Figure 4] show a marked increase in performance of the system, especially in the case of Lichen Planus which underperformed in the MaxEnt model. Non-linear models like ANN and DT learns the underlying pattern and gives better accuracy.

VII. RESULTS AND DISCUSSIONS

The novel method of using a dual stage system has given very promising results in identification of skin diseases with accuracies of up to 95%. Comparison of our work with related works in this domain has revealed stark differences in the implementation and performance. None of the existing solutions for identification of skin diseases handle the six diseases that we have proposed in this paper. One of the

papers in this domain [1], proposes usage of data mining for detection although this data set is unavailable. Reference [2] uses only image processing and some clinical attributes in order to identify the type of skin disease, which we found to be inadequate due to the absence of histopathological attributes.

VIII. COMBINED RESULTS

The mobile application developed on the principles as described above produce better results than any application in this space owing to the two stage refinement in detection. Though each stage in itself produces fairly accurate results, combining the two stages increases the accuracy, making this application an efficient and dependable system for dermatological disease detection. Furthermore, this can be used as a reliable real time teaching tool for medical students in the dermatology stream. As an added advantage, this application can also be used by the common user as we have been able to achieve fairly accurate detection rate by Computer Vision techniques alone.

IX. FUTURE WORK

The system suffers from inaccuracies when it is tasked with detection of diseases on varying skin colors. As part of our future work, we would like to make the system develop immunity to the varying skin colors. Our focus in this system has been on the six of the most common dermatological diseases. We intend to continue working with the doctors to come up with better feature functions in order to broaden the number of diseases that the system can detect.

ACKNOWLEDGEMENT

We would like to thank Prof. Channabasavanna Bankapur from Computer Science Department of P.E.S Institute of Technology for his continued guidance and support without whom this project would have been difficult to finish.

REFERENCES

[1]"Expert System for Diagnosis of Skin Diseases", *International Journal of Science and Technology*, vol. 4, no. 1, 2015.

[2]R. Yasir, M. Rahman and N. Ahmed, "Dermatological Disease Detection using Image Processing and Artificial Neural Network".

[3]R. Parikh and D. Shah, "A Survey on Computer Vision Based Diagnosis for Skin Lesion Detection", *International Journal of Engineering Science and Innovative Technology*, vol. 2, no. 2, 2013.

[4]K. Gimpel and N. Smith, "Softmax-Margin CRFs: Training Log-Linear Models with Cost Functions".

[5]A. Narayana, "Decision Trees", Bangalore, 2015.

[6]N. Fujishima and K. Hoshino, "Fingernail Detection Method from Hand Images including Palm", in *IAPR International Conference on Machine Vision Applications*, Kyoto,Japan, 2013.

[7] Abbadi, "Psoriasis Detection Using Skin Color and Texture Features", *Journal of Computer Science*, vol. 6, no. 6, pp. 648-652, 2010.

[8]B. Dhandra, S. Soma, S. Reddy and G. Mukarambi, "Color Histogram Approach for Analysis of Psoriasis Skin Disease", in *Int. Conf. on Multimedia Processing*.

[9]A. Mittra and D. Parekh, "Automated Detection of Skin Diseases Using Texture Features", *International Journal of Engineering Science and Technology*, vol. 3, no. 6, pp. 4801-4808, 2011.

[10]R. Bhardwaj and S. Vatta, "Implementation of ID3 Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 6, pp. 845-851, 2013.

[11]"Unsupervised Feature Learning and Deep Learning Tutorial", *Ufldl.stanford.edu*, 2016. [Online]. Available: http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/. [Accessed: 21- Feb- 2016].

[12]R. Bansal and M. Saini, "A Method for Automatic Skin Cancer Detection", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 9, no. 5, pp. 529-533, 2015.

[13]"Diagnosing Skin Diseases Using an Artificial Neural Network", in *International Conference on Adaptive Science & Technology*, 2009, pp. 187-191.

[14] M. Arifin, M. Kibria, A. Firoze and M. Amin, "Dermatological Disease Diagnosis Using Colour Skin Images", in *International Conference on Machine Learning and Cybernetics*, 2012.

[15] S. Abdul-Rahman, M. Yusoff, A. Mohamed and S. Mutalib, "Dermatology Diagnosis with Feature Selection Methods and Artificial Neural Network", in *IEEE EMBSInternational Conference on Biomedical Engineering and Sciences*, 2012.

[16]M. Asghar, M. Asghar, S. Saqib and B. Ahmad, "Diagnosis of Skin Diseases using Online Expert System", *International Journal of Computer Science and Information Security*, vol. 9, no. 6, pp. 323-325, 2011.

[17]T. Mitchell, *Machine Learning*. McGraw-Hill Science, 1997, pp. 52-126.

[18] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[19] G. Cybenko, "Approximation by superpositions of a sigmoidal function", *Mathematics of Control, Signals, and Systems*, vol. 5, no. 4, pp. 455-455, 1992.

[20]F. Tushabe, E. Mwebaze and F.Kiwanuka, "An image-based diagnosis of virus and bacterialskin infections", in *The International Conference on Complications in Interventional Radiology*, 2011.

[21]D. Okuboyejo, O. Olugbara, and S. Odunaike, "Automating Skin Disease Diagnosis Using Image Classification", in *World Congress on Engineering and Computer Science*, 2013.

[22]T. Lee, V. Ng, R. Ghalleger and A. Coldman, "DullRazor: a software approach to hair removal from images", *Computation Bio Med*, vol. 6, no. 27, pp. 533-544, 1997.