IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

Prediction of Skin Diseases using Data Mining Techniques

S. Reena Parvin[1], O.A. Mohamed Jafar[2]

Assistant Professor, Computer Science, Syed Ammal Arts and Science, Ramanathapuram, Tamil Nadu[1]

Assoc. Professor, PG &Research Dept of Computer Science, Jamal Mohamed College, Tiruchirappalli, Tamil Nadu[2]

Abstract: Skin Disease prediction has become important in a variety of applications such as health insurance, tailored health communication and public health. Due to the costs for dermatologists to monitor every patient, there is a need for an computerized system to evaluate a patient‘s risk of melanoma using images of their skin lesions captured using a standard digital camera. The traditional diagnosis technique aims at improving the quality of existing diagnostic systems by proposing advanced feature extraction and classification methods. In the Proposed method, 40 digital images collected from AOCD unit database and another 40 digital images from MIT unit database. These images are subjected to pre-processing the images using Gaussian Filter technique. Then these images are undergone image segmentation using K-means clustering algorithm to partitioning the disease affected area and non-affected area. Feature extraction is performed using Grey Level Co-occurrence Matrix (GLCM) for examining texture which gave the statistical parameters, for better classification efficiency. The multi-SVM (Support Vector Machine) classifier is supervised learning models with associated algorithms that analyze database images for classification analysis. The diagnosis system involves two stages of process such as training and testing, the Features values of the training data set are compared to the testing data set of each type. Finally the performance analysis compared three algorithms such as Multi-SVM classifier, K-NN and Naïve Bayesian classifier. The overall accuracy of using Multi-SVM classifier is 97% to 98%.

Keywords: Skin disease prediction, GLCM, Gaussian Filter, Multi SVM, K-NN, Naïve Bayesian, K-means.

I. INTRODUCTION

Skin is the largest organ is our body. Skin is that, which helps to protects the body from infection, heat, injury, and any type of damage which is caused by ultraviolet (UV) radiation. It helps to make vitamin D. So production of skin from diseases is the significant and intricate job in medicine.

Skin does many different things,
i) Covers the internal organs and helps protect them from injury
ii) Serves as a barrier to germs such as bacteria
iii) Prevents the loss of too much water and other fluids
iv) Helps control body temperature
v) To make the vitamin D.

So protect our body from skin diseases is must to lead a healthy life. The main aim of this study is to perform different types of skin disease images are group together in data set and pre-processing the input image using Gaussian filter and apply separate the subgroups affected area and non affected area using k-means clustering algorithm and apply feature extraction using GLCM method and find the diseases. Image processing operations can be roughly divided into three major categories, Image Compression, Image Enhancement and Restoration, and Measurement Extraction. It involves reducing the amount of memory needed to store a digital image. Image defects which could be caused by the digitization process or by faults in the imaging set-up (for example, bad lighting) can be corrected using Image Enhancement techniques.

II. LITERATURE REVIEW

Mugdha S Manerkar et al.,[1] using C means and Watershed algorithms for image segmentation. Feature extraction is performed using Gray Level Co-ocurrence Matrix (GLCM) and Image Quality Assessment (IQA) methods for texture which gave the statistical parameters of each algorithm, Features values of training data and testing. G.Ramya and J Rajeshkumar [2] used GLCM method for extracting features from the segmented diseased and classified the skin cancers based on fuzzy classification, higher accuracy compared to existing one. B.Gohila vani et al.,[3] used a novel

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

texture based skin lesion segmentation algorithm to classify stages of cancer by Probabilistic Neural Network(PNN) on basis of learning and training samples of data. Kawsar Ahmed et al.,[4] used pre-processing data is clustered using k-means clustering algorithm for separating relevant and non relevant data to skin cancer. Frequent patterns are discovered using MAFIA algorithm. AprioriTid and decision tree algorithms of extracting the frequent patterns from clustered dataset. I.Vijaya et al.,[5] focus on non-melanoma skin cancers and classify the types, predict the type of disease accurately using support vector machine. Color rate and texture features are extracted preprocessed training dataset. Y.P.Gowaramma et al., [6] used marker controlled watershed segmentation method k-nn classifier along with curvelet filter. J.Priyadharshini et al., [7] using classification algorithms to evaluate the dermatological diseases in various dimensions. Naive bayes algorithm technique for constructing classifiers. E.Barati et al., [8] emphasized highlighted and provided the various mining methodologies, they also found that apriori algorithm is best for feature extraction. Discussion result classification technique applied for good results. A.A.L.C Amarathunga et al.,[9] predict the climate and living situations skin diseases. Image of skin disease captured to eliminate noises, pre-processing done by Gaussian smoothing process. Seperate the region of the disease using image segmentation algorithm. Feature extraction carriedout using classification algorithm. Madhura Rambhajani et al.,[10] using Bayesian techniques with best first search feature selection used in order to classify the dermatology diseases algorithm.

III.METHODOLOGY

The digital skin disease images were taken as from AOCD, MIT datasets and pre-processing techniques were applied to these input images. K-means algorithm was applied on pre-processed images to segment the skin diseases automatically.

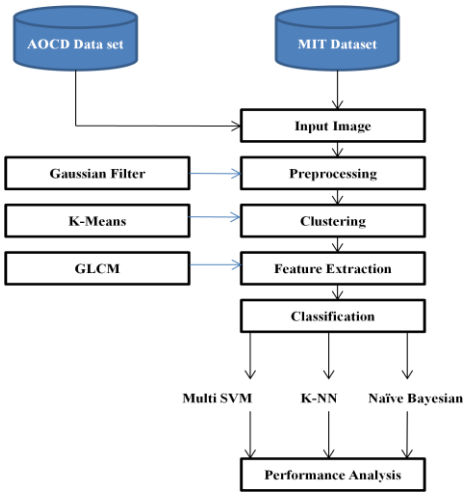

Fig. 1 Data Flow Diagram

A. Input Image

The input images for this project are the digital images of various skin diseases. These images were taken as input from the data set.

B. Pre-processing

In pre-processing apply Gaussian filtering to our input image. Gaussian filtering is often used to remove the noise from the image. Here we used wiener function to our input image. **Gaussian filter** is windowed filter of linear class, by its nature is weighted mean. Named after famous scientist Carl Gauss because weights in the filter calculated according to Gaussian distribution.

The Gaussian Smoothing Operator performs a weighted average of surrounding pixels based on the Gaussian distribution. It is used to remove Gaussian noise and is a realistic model of defocused lens. Sigma defines the amount of blurring. The radius slider is used to control how large the template is. Large values for sigma will only give large blurring for larger template sizes. Noise can be added using the sliders.

C. Clustering

Clustering is a process of separating dataset into subgroups according to the unique feature. Clustering separated the dataset into relevant and non-relevant dataset. Clustering is another tedious term of data mining. The clustering problem has been addressed in numerous contents besides being proven beneficial in many applications. The goal of clustering is to classify objects or data into a number of categories or classes where each class contains identical

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

feature. The main benefits of clustering are that the data object is assigned to an unknown class that have unique feature and reduces the memory.
The k-means clustering is a widely recognized clustering tool that is used for robotics, diseases and artificial intelligence application purposes. Here k is a positive integer representing the number of clusters. The pre-processed data is clustered using the k-means clustering algorithm with the value of k equal to 2. This represents there is two clusters where one is relevant data and another contains non-relevant data.

D. Feature Extraction
The features were extracted from the images based on Gray Level Co-occurrence matrix (GLCM). GLCM statistics like energy, correlation, entropy, and homogeneity were extracted from the center tumor and the whole skin region. The process calculation of the GLCM statistics was applied in different distances and in different angles. In each case of the calculation of the GLCM statistics min, mean, max, and difference (i.e., max-min) values were extracted from the image regions. The extracted values denote the texture of the input images in an efficient manner. The values extracted from the images were then saved as the features. Few of the common statistics applied to co-occurrence probabilities are given the following table with related formulas.

E. Classification
Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Classification divides data samples into target classes. The classification technique predicts the target class for each data points.

F. Multi-SVM Classifier
This defines a grouping of all the classes in two disjoint groups of classes. This grouping is then used to train a SVM classifier in the root node of the decision tree, using the samples of the first group as positive examples and the samples of the second group as negative examples. The classes from the first clustering group are being assigned to the first (left) subtree, while the classes of the second clustering group are being assigned to the (right) second subtree. The process continues recursively until there is only one class per group which defines a leaf in the decision tree.

G. K-NN classifier
Classification has been done by KNN classifier. The k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. The input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

H. Naïve Bayesian Classifier
In machine learning, naïve Bayesian classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong (naive) independence assumptions between the features. Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

IV. PERFORMANCE ANALYSIS

The accuracy, sensitivity and specificity of the classifier are measured. The accuracy represents the efficiency of the process. The sensitivity shows how the algorithm gives correct classification. The specificity shows how the algorithm rejects the wrongly classification results. We designed a spatial consistency constraint in a graphical model to improve the detection performance. The performance of the process is measured based on the calculation of Accuracy, Area under curve of the process.

V. EXPERIMENTAL RESULTS

The above stated algorithms are implemented in MATLAB. Around 40 digital skin diseases images are taken from two different data sets MIT and AOCD these are subjected to pre-processing techniques such as image resizing, image format conversion, contrast enhancement.
The pre-processed images are given as inputs for smoothened using Gaussian filter for better accuracy images partitioned into affected area and non-affected area using k-means clustering algorithm. Features are extracted by

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

GLCM method. Features of test images are compared with the training data set by the disease classification into specific categories using multi SVM classifier. In Experimental results the performance measure of AOCD data set are presented in Fig.3 and MIT dataset are presented in Fig.5 that representing the performance measure of SVM, k-NN, Naïve Bayesian classifiers. Overall accuracy of this study ranges from 97% to 98%.

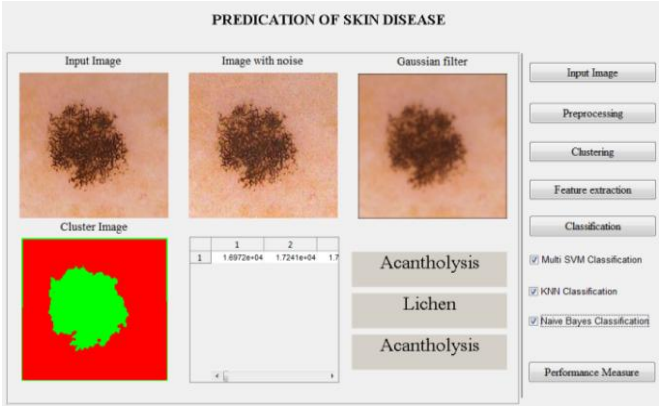

Fig.2 . The GUI Form of Skin Disease Prediction AOCD dataset

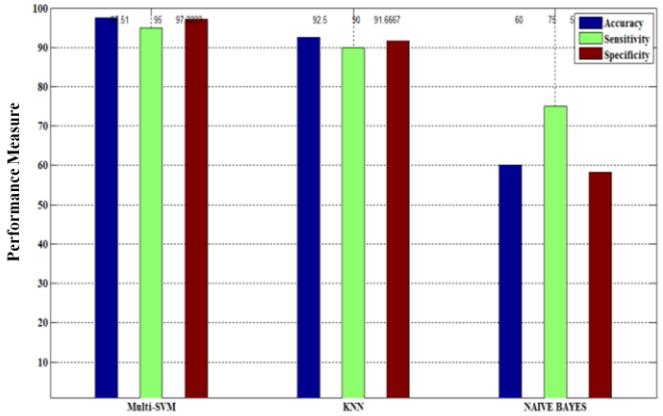

Fig. 3. Performance measure of Dataset1

TABLE I **THE PERFORMANCE OF ACOD DATA SET**

Classification Algorithms	The Performance of ACOD dataset		
	Accuracy	Sensitivity	Specificity
Multi-SVM	97.5	95	97.2
K-NN	90	70	91.4
Naive Bayesian	60	75	58.3

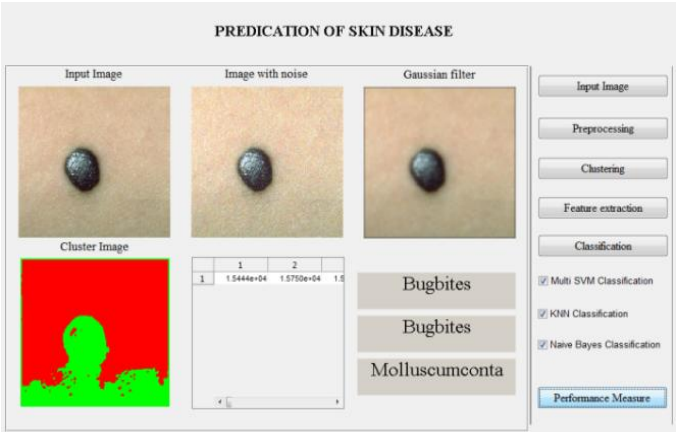

Fig.4 . The GUI Form of Skin Disease Prediction MIT dataset

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

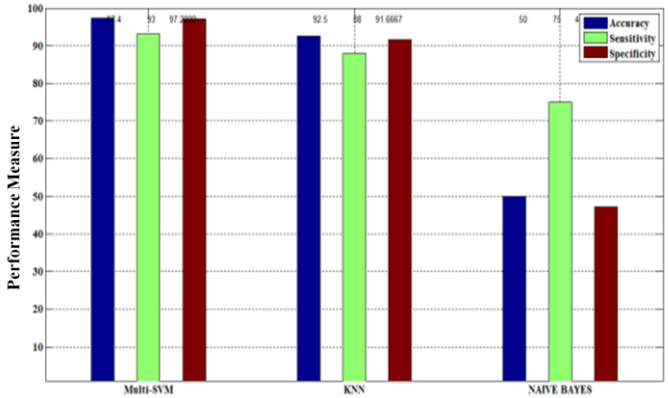

Fig. 5. Performance measure of Dataset2

TABLE III **THE PERFORMANCE OF MIT DATA SET**

Classification Algorithms	The Performance of MIT dataset		
	Accuracy	Sensitivity	Specificity
Multi-SVM	97.4	93	97.2
K-NN	90	68	91.4
Naive Bayesian	55	75	52

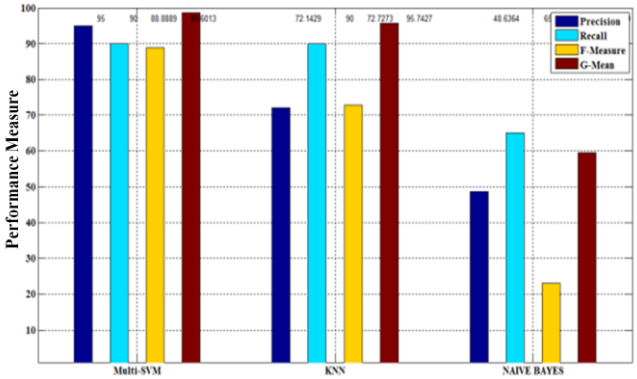

Fig. 5. Performance measures

TABLE IIIII **THE PERFORMANCE MEASURES**

Classification Algorithms	The Performance Measures			
	Precision	Recall	F-Measures	G-Mean
Multi-SVM	95	90	88.8	98.6
K-NN	72.1	70	66.6	85.5
Naive Bayesian	51.6	65	27.2	66.1

VI. CONCLUSION

In proposed system we compare one different established feature. We also experiment with different classifier architectures. The feature sets and most of the classifier architectures we tested, provide a similar performance. To improve the performance the results of the experiments indicate that this method has much faster training and testing times than the widely used multi SVM methods. Support vector machine has been used to train the model. The accuracy, sensitivity and specificity of the classifier are measured. The outcome of the experiments proves that support vector machine is effectual. As future enhancement of this research work, more features that help to increase the classification accuracy can be identified, extracted and used for learning. In future to implement multi SVM classifier

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

with some other extracting features. A system framework is presented to recognize multiple kinds of activities from videos by an SVM multi-class classifier with binary tree architecture. The thought of hierarchical classification is introduced and multiple SVMs are aggregated to accomplish the recognition of actions. Each SVM in the multi-class classifier is trained separately to achieve its best classification performance by choosing proper features before they are aggregated.

REFERENCES

[1] Mugdha Smanerkar, shashwata harsh, juhi saxena, simanta p sarma, dr. u. snekhalatha, dr.m. anburajan, "classification of skin disease using multi svm classifier" 3rd international conference on electrical, electronics, engineering trends, communication, optimization and sciences (eeecos)-2016.
[2] G.Ramya, J.Rajeshkumar "Novel method for segmentation of skin lesions from digital images", international research journal of engineering and technology vol:02 issue:08 November 2015.
[3] B.Gohila vani et al. "Segmentation and Classification of Skin Lesions Based on Texture Features" Int. Journal of Engineering Research and Applications www.ijera.com ISSN : 2248-9622, Vol. 4, Issue 12(Part 6), December 2014, pp.197-203.
[4] Kawsar Ahmed, Tasnuba Jesmin, Md. Zamilur Rahman "Early Prevention and Detection of Skin Cancer Risk using Data Mining" International Journal of Computer Applications (0975 – 8887) Volume 62– No.4, January 2013.
[5] I.Vijaya M. S "Categorization of Non-Melanoma Skin Lesion Diseases Using Support Vector Machine and Its Variants". International Journal of Medical Imaging. Vol. 3, No. 2, 2015, pp. 34-40. doi: 10.11648/j.ijmi.20150302.15
[6] Y.P.Gowaramma et al., used marker controlled watershed segmentation method k-nn classifier along with curvelet filter.
[7] J. Priyadharshini "A Classification via Clustering Approach for Enhancing the Prediction Accuracy of Erythemato-squamous (Dermatology) Diseases" IJSRD - International Journal for Scientific Research & Development| Vol. 3, Issue 06, 2015 | ISSN (online): 2321-0613.
[8] E.Barati et al., "A survey on utilization of data mining approach for dermatological skin diseases prediction" Journal of selected areas in health informatics march 2011.
[9] A.A.L.C. Amarathunga, et al.,"Expert system for diagnosis of skin diseases" International journal of scientific & technology research volume 4, issue 01, january 2015 issn 2277-8616 174 ijstr©2015.
[10] MadhuraRambhajani "Classification of Dermatology Diseases through Bayes net and Best First Search" International Journal of Advanced Research in Computer and Communication Engineering" Vol. 4, Issue 5, May 2015.