Prediction of systemic lupus erythematosus based on cutaneous manifestations using machine learning models

S Samundeswari, V Ramalingam, B Latha and S Palanivel

Abstract
Systemic Lupus Erythematosus (SLE) or Lupus is an immune system illness in which the human immune system becomes hyperactive and affects all tissues. Recognizing SLE is challenging, and the doctor may take a long time to diagnose this intricate disease accurately. This experimental study employed the use of images gathered from various hospitals in Tamilnadu. Images of 400 patients (200 SLE and 200 normal) were analyzed in this study. Machine learning models can offer a helpful tool to the doctor to predict the disease at an early stage before it becomes severe. Three classifiers, Naïve Bayes, Multilayer perceptron (MLP), and Random Forest are considered to classify the patients. Experimental results demonstrate that the MLP gives a higher accuracy than other models for recognizing SLE.

Keywords: systemic lupus erythematosus (SLE), cutaneous manifestation, multilayer perceptron (MLP), naïve bayes (NB), random forest (RF), lupus

1. Introduction
Systemic Lupus Erythematosus (SLE) or Lupus is an immune system illness in which the human immune system becomes hyperactive and affects all tissues. Systemic Lupus Erythematosus (SLE) is the immune system illness in which body tissues are affected by its own particular immune system. SLE can influence various organs in the body, like heart, lungs, liver, kidney, and central nervous system. At the point when confronted with men, the infection seems nine times ordinarily in ladies; particularly it is more typical in ladies of younger bearing ages. Sadly, the reason for lupus is flighty with times of ailment and health. This variety prompts diagnosis of SLE exceptionally difficult.

The greater part of the lupus patients have cutaneous signs. According to Gilliam classification, Lupus skin lesions are classified into lupus-specific (e.g., malar rash) and non-specific skin lesions (e.g., alopecia) [1]. Lupus specific is grouped into three noteworthy subtypes: systemic or acute lupus erythematosus (ACLE), subacute cutaneous lupus erythematosus (SCLE) and chronic cutaneous lupus erythematosus (CCLE) [1-3]. Culture, Environment and hereditary qualities make fluctuation in frequency and infection seriousness among different racial and ethnic groups [5]. Laman et al [6] proposed the variance in the skin lesions from malar rash, discoid lupus to oral ulcer, alopecia, bullous lesions and so forth. The American College of Rheumatology (ACR) proposed 11 criteria as a characterization tool for recognizing the SLE in the patients [4]. Cutaneous lesions ought to fulfill four of the 11 re-examined ACR criteria of SLE. Lupus-specific skin lesions can be effectively analyzed while lupus non-specific skin lesions are related with the failure of different organs and subsequently patients should be continuously monitored [7] for the severity of the disease. Therefore, a rheumatologist ought to exclusively assess the patients in view of the clinical and symptomatic criteria for viable diagnosis and treatment.

Color Histogram is broadly used to extract color features of an image [8]. It is used to discover the frequency distribution of the color bins in an image. Pixels having same intensity are counted and stored. As color histogram can be easily calculated, it is vital for analyzing the images [10]. Nanni et al [11] proposed combination of different surface descriptors utilizing GLCM and have been tried for various classification of medical images. Gabor filters, local binary patterns (LBP), scale-invariant feature transform (SIFT), histogram of oriented gradients (HOG), speeded up robust feature (SURF) techniques are also used to extract the texture features. The Gray - Level Co-Occurrence Matrix (GLCM) feature descriptors are
extracted by finding out the relationship between the two neighbouring pixels [9].

2. Materials and Methods

Patients who fulfilled the revised 2012 SLICC ACR characterization criteria for SLE are selected and their images are collected from different hospitals in Tami Nadu. A sample of 400 images (200 SLE and 200 Non-SLE) is taken for this study. These 400 image samples are used to classify the images into SLE and Non-SLE. Some of the cutaneous manifestations of SLE images are shown in figure 1. Pre-processing is the first step in the analysis of the images. The undesirable noise present in the images should be removed so as to enhance the effectiveness of our framework. Hence Medium filter, a nonlinear digital filtering technique is used to remove the undesirable noise.

Fig 1: Image samples of cutaneous manifestation of SLE patients

2.1. Methodology

The objective of the paper is to analyze and compare the performance of the classifiers Naïve Bayes, MLP and Random Forest by extracting the color histogram and texture features in HSV color space from the various images. The general methodology overview of the proposed approach is outlined in figure 2. The proposed framework comprises of three phases: (i) Pre-processing, to remove the noise using Median filtering (ii) Feature extraction to extract the color histogram and texture features in HSV color space (iii) classifiers, Naïve Bayes, MLP and Random Forest to classify the images into SLE and Non-SLE. The performances of the classifiers are assessed by computing the accuracy, sensitivity and specificity.

Fig 2: Proposed methodology to classify the images.

2.2 Feature Extraction

Feature extraction is one of the important steps for analysis process. This technique is used to extract a set of highly informative features from the images. In this proposed work, color histogram based features and the texture based features are used to classify the images. The color histogram features of mean, standard deviation, skewness, kurtosis, and energy are computed based on the probability distribution of the histogram intensity levels. 5 histogram features are extracted for each channel of HSV color space. The Gray-Level Co-Occurrence Matrix (GLCM) method is used to extract the texture based features. The GLCM function calculates the spatial association of the pair of pixels with specific values and creates a co-occurrence matrix. The texture features are extracted from this matrix. The images are decomposed into three channels of HSV color space thereby obtaining three different images to extract the texture features of the color images. Then the features are extracted for each channel of HSV color space. In our proposed work, 8 features namely, autocorrelation, contrast, cluster prominence, dissimilarity, energy, entropy, sum average and sum variance are computed. Hence a total of 39 features (5 Histogram features and 8 texture features for each channel of the color space) are extracted from the images and given as an input to the classifiers. The extracted features are shown in Table 1.

<table>
<thead>
<tr>
<th>Histogram based features</th>
<th>Texture features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Autocorrelation</td>
</tr>
<tr>
<td>Variance</td>
<td>Contrast</td>
</tr>
<tr>
<td>Skewness</td>
<td>Cluster Prominence</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>Dissimilarity</td>
</tr>
<tr>
<td>Energy</td>
<td>Energy</td>
</tr>
<tr>
<td>Entropy</td>
<td>Sum average</td>
</tr>
<tr>
<td></td>
<td>Sum variance</td>
</tr>
</tbody>
</table>

2.3 Classification Techniques

Classifiers, Naïve Bayes, Multilayer Perceptron and Random Forest are used to classify the images using the features descriptors. Naïve Bayes classifier is simple classifier which is based on Bayes Theorem of conditional probability and strong independence assumptions. This classifier emphasizes on measure of probability that whether the document A belongs to class B or not. It is based on the assumption that the occurrence or non occurrence of a particular attribute is unrelated to the occurrence or non occurrence of a particular attribute [14]. MLP neural networks [15] comprise of three layers of nodes. It has (i) input layer, (ii) output layer, and (iii) hidden layer and each of which contains input node (s) otherwise called as sensory nodes, output node (s) called as responding nodes, and hidden node (s), respectively. The hidden and output nodes are the processing elements which contains the activation function that is used to obtain the output. Random Forest [15] uses decision tree as base classifier. Random Forest generates multiple decision trees; the randomization is present in two ways: (1) random sampling of data for bootstrap samples as it is done in bagging and (2) random selection of input features for generating individual base decision trees. Strength of individual decision tree classifier and correlation among base trees are key issues which decide generalization error of a Random Forest classifier.

3. Experimental Results

In this proposed work, images of both SLE and normal patients are collected. These images are used to analyze and evaluate the performance of the classifiers for identifying the lupus. Feature descriptors are extracted from the images based
on GLCM and color histogram in HSV color space. The classifiers, Naïve Bayes, MLP and Random Forest are trained and tested to produce the required output for the given features. Precision, recall, specificity and accuracy are the performance indicators used to evaluate the performance of the above three classifiers for the classification of the images into SLE and Non-SLE.

Precision = $\frac{TP}{TP + FP}$ (1)

Recall = $\frac{TP}{TP + FN}$ (2)

Specificity = $\frac{TN}{TN + FP}$ (3)

Accuracy = $\frac{(TP + TN)/ (TP + TN + FP + FN)}{4}$

Where TP (True Positive) represents the number of correctly classified SLE images, FP (False Positive) represents the number of misclassified SLE images, FN (False Negative) represents the number of misclassified Non-SLE images, and TN (True Negative) represents the number of correctly classified Non-SLE images.

3.1 Performance Evaluation

Table 2 outlines the classifiers with corresponding precision and recall values. Figure 3 is the graphical representation of Table 2. Figure 3 shows that MLP performed better than the other classifiers Naïve Bayes and Random Forest.

### Table 2: Classifiers with corresponding Precision and Recall values

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLE</td>
<td>NON-SLE</td>
<td>SLE</td>
</tr>
<tr>
<td>NB</td>
<td>75.4</td>
<td>76.1</td>
</tr>
<tr>
<td>MLP</td>
<td>0.857</td>
<td>0.868</td>
</tr>
<tr>
<td>RF</td>
<td>82.3</td>
<td>85.3</td>
</tr>
</tbody>
</table>

The performance measures (sensitivity, specificity and accuracy) of various classifiers are shown in Table 3. From Table 3, one can infer that again MLP outperforms Naïve Bayes and Random Forest in all the performance descriptors, namely, sensitivity (87%), specificity (85.5%) and accuracy (86.25%).

### Table 3: Classifiers versus various performance measures

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy Value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>76.50</td>
<td>75.00</td>
<td>75.75</td>
</tr>
<tr>
<td>MLP</td>
<td>87.00</td>
<td>85.50</td>
<td>86.25</td>
</tr>
<tr>
<td>SVM</td>
<td>86.00</td>
<td>81.50</td>
<td>83.75</td>
</tr>
</tbody>
</table>

Figure 4 shows the graphical representation of Table 3. Naïve Bayes classifier shows poor performance when compared with MLP and Random Forest.

4. Conclusion

Prediction of Systemic Lupus Erythematosus based on cutaneous manifestations is proposed. Cutaneous manifestations give imperative information to the rheumatologists to analyze whether the patient has lupus or not. As Lupus affects all the major parts of the body such as lungs, heart, kidneys and central nervous system, it is very important to diagnose the lupus at the early stage to ensure speedy treatment. The diagnosing methodology use different classifier algorithms for the classification of SLE from a group of patients. MLP shows the maximum performance than other classifiers Naïve Bayes and Random Forest. MLP yields maximum of 86.25% as accuracy. If precision, recall, accuracy are considered together, one can conclude that MLP is the better choice than other classifiers, namely, Naïve Bayes and Random Forest. Further this decision is supported by Rheumatologist also.

5. References