Systemic Lupus Erythematosus manifestation using ID3 Algorithm – A clinical Analysis

S. Gomathi[#], Dr. V. Narayani[*]

[#]*Department of Information and Computer Technologies, Sri Krishna Arts and Science College*

**Department of MCA, Karpagam College of Engineering*
Coimbatore

[1]gomathisrinivasan88@gmail.com

[2]narayaniv79@rediffmail.com

***Abstract*— Discovering hidden patterns in medical data and relationship between them is often fallow. Classification technique in data mining is used to discover the hidden knowledge from enormous data. This work is done on predicting the risk of Systemic Lupus Erythematosus (SLE)/ Lupus using data mining classification technique. Decision tree algorithm is used for training set of data. A new proposed framework and an enhanced algorithm is proposed. The classification algorithm is used to reduce the complexity and to increase the performance.**

***Keywords*—– ID3, Re-enactment, malar rash, SLEDAI, antinuclear antibody, lupus**

I. Introduction

Systemic Lupus Erythematosus (SLE) is an autoimmune, multi system disease which can affect any system of the human body including central nervous system which include cognitive and mood disorder, anxiety, depression and psychosis [8]. SLE will develop distinct immunologic abnormalities particularly antinuclear, antiphospolipid [6]. There are four main factors involved in affecting lupus (i) genetic, (ii) hormonal, (iii) immunologic and (iv) environmental [4, 9]. This disease found difficult to diagnose and cannot be predicted with single parameter. It can be identified with a combination of laboratory and clinical criteria [7]. American college of rheumatology established 11 clinical and laboratory criteria among which 4 criteria must be satisfied to diagnose and treat SLE [10, 11].

The two primary goals of data mining is descriptive and prediction. Description is about finding the patterns to describe data which can be interpreted by human whereas prediction involves fields or variables to predict future unknown values of other variables [5]. Predicting disease is vital role of data mining. Many disease like diabetes, lung cancer, breast cancer, thyroid etc., are predicted using data mining classification techniques. The modern application of data mining include in predicting SLE disease. A new model is developed using decision tree algorithm. The hidden patterns of patients are utilized for clinical diagnosis for widely distributed unexploited medical data which will be further converted into organized form.

978-1-4799-4674-7/14/$31.00©2014 IEEE

II. Literature Review

Sayed et.al., [1] designed a decision support system to predict heart disease using ID3 algorithm & multilayer perceptron with back propagation as training algorithm. The work was done to predict the risk of heart attack. It covers the main objective to utilize the knowledge of previous history about the patient.

S. Vijayarani et.al., [2] surveyed about data mining techniques to predict the various type of diseases. These techniques are extremely applied for all the fields like health care, banking agriculture etc.,

Vikas chaurasia et.al., [3] used three popular data mining classification technique CART, ID3 & C4.5. The research shows the accuracy of ID3. Training and simulator error evaluation was also done. The model was implemented using WEKA tool.

Renu Saigal et.al., [4] shows that survival rate of SLE are approximately 80% in 10 years after diagnosis and approximately 65% in 20 years. A study conducted on Zimbabwe showed that renal involvement was common than photosensitivity and serositis which was less common in United States. Based on the study the author concluded that the wide variation in the natural history of lupus is based on geographical groups and different ethnic. The 3 stages of lupus are mild, moderate and severe.

III. Problem Specification

Prediction of the lupus disease needs a high data storage and good decision system, since the disease may affect any part of the body and which is not common to all patients. It needs a good and effective method to predict earlier. Thus this paper suggested a new methodology to predict and analyse the disease.

IV. Proposed Work

The proposed work is to design a framework, to create an algorithm and the attributes which are involved in predicting the disease.

A. *Proposed Framework*

Fig 1 depicts the proposed framework which is divided into 3 phases.

Phase I: The first phase is acquiring phase which is acquiring the data from the lupus care specialist. The dataset about the existing lupus patient along with their complete history will be collected and in turn stored in database.

Phase II: Data refinement phase is refining the necessary data from the database. This phase involves five steps (i) data pre processing: where the data will be processed, (ii) Data re-enactment: where the data will be further processed and the necessary data will be gathered. (iii) dataset classification: the data set is further classified based on plasma and serum blood count, (iv) access activity and severity: the disease is scored based on the WHO classification which shows how severe the disease affected the patient and (v) Monitor disease activity: which is to monitor how the disease is spreading to other parts of the body.

Phase III: Classification and prediction phase is where the ID3 algorithm is actually applied. The algorithm will be generated based on the SLEEDAI score, ACR diagnosis criteria. Finally the validation process will be carried out.

B. *ID3 Algorithm*

ID3 is a predictive model which was invented by Ross Quinlan in 1979. ID3 is used to build decision tree with information theory which was invented in 1948 [5]. It is a top down approach without backtracking. To select the best attribute for classification, information gain is used. Entropy is used to measure the uncertainty which ranges from 0 to 1.

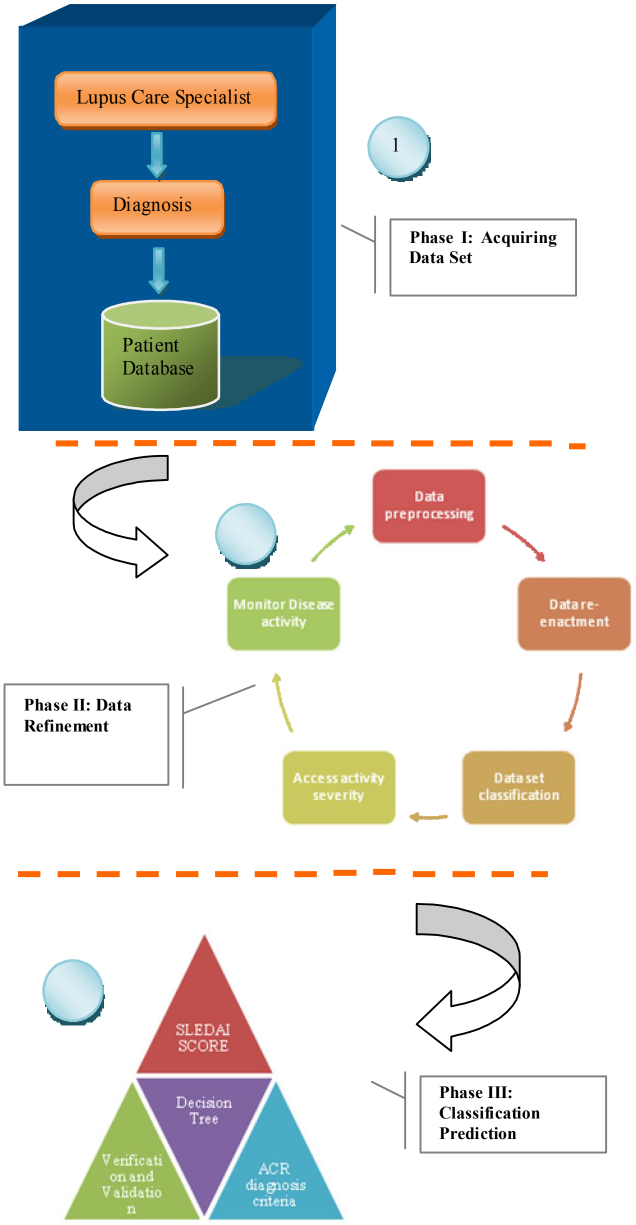

Fig. 1 Modern framework to predict Lupus disease

Information gain used to measure the expected reduction in entropy. Initially ID3 calculates the gain of all attributes finally selects the one with the highest gain. The attribute with highest information gain will be located as root node in decision tree.

```
Step 1: Compute classification entropy
Step 2: Calculate information gain for each attribute
Step 3: Acquire attribute with highest information gain
Step 4: Remove node attribute for further calculations
Step 5: Loop 2-3 till attribute have been used
Algo ID3(input attribute, output attribute, training data)
{
        If(training data = 0)
        {
            Return (single node= failure);
        }
        If(records = positive)
        {
            Return(single node = positive value);
        }
        If(records = negative)
        {
            Return (single node = negative value);
        }
        If(input attribute = 0)
        {
            Return(single node = most recent value);
        }
        Else
        {
            Calculate information gain for each attribute;
            Divide the attribute with highest information gain value;
            Return tree with root node A and arcs A1,A2...Am;
            Call algoID3 repeatedly till all attributes have been used;
        }
}
```

Fig.2 ID3 Algorithm

C. Important attributes for SLE

Table I depicted the 9 important attributes to predict SLE disease. The table is divided into three columns, ID, attribute name and attribute domain values. Based on the data acquired, the decision tree will be generated. SLEDAI is disease Activity Index score which is used to show the severity of the disease which is shown in Table II. The decision tree shows the final prediction. 20 patients data has been given as input to the ID3 algorithm. Table III shows the analysis result and the clinical profile is shown in Table IV. Fig 2 shows the algorithm and Fig 3 shows the experimental result of the algorithm.

TABLE I
FONT SIZES FOR PAPERS

ID	Attributes name	Domain Values
1	Age (in years)	1: <=20 2: >=21 and <=30 3: >=31 and <=40 4: >=41 and <=50 5: >51
2	Gender	0: female 1: male
3	Sample type	0: Serum 1: Plasma 2: Urine
4	Ethnicity	0: African 1: American 2: Caucasian 3: Hispanic 4: Asian
5	Disease activity	0: Mild/Flare 1: Moderate/Chronic 2: Severe/Log quisence
6	ACR criteria	1: malar rash 2: discoid rash 3: photosensitivity 4: oral ulcers 5: non erosive arthritis 6: pleuritis 7: renal disorders 8: neurologic disorder 9: hematologic disorder 10: immunologic disorders 11: antinuclear antibody
7	Organs involved	0: none 1: Skin 2: Joints 3: Musculoskeletal 4: Blood 5: Brain 6: Lung 7: CNS 8: Vascular 9: eyes 10: heart 11: pulmunory 12: gastrointensional 13: Moouth 15: extremities
8	Tests	0: ANA 1: CBC 2: Chest X-ray 3: Kidney biopsy 4: Urinalysis 5: Rheumatoid test facts 6:Liver function blood test 7: ESR

TABLE III
SLEEDAI SCORE

Symptoms	SLEEDAI Score
Seizure	8
Psychosis	8
Organic brain syndrome	8
Visual disturbance	8
Cranial nerve disorder	8
Lupus headache	8
Cerebrovascular	8
Vasculitis	8
Arthritis	4
Myositis	4
Urinary casts	4
Hematuria	4
Protenuria	4
Pyuria	4
New rash	4
Alopecia	2
Mucosal ulcers	2
Pleurisy	2
Pericarditis	2
Low complement	2
Increased DNA binding	2
Fever	1
Thrombocytopenia	1
Leukopenia	1

D. Experimental Result

The performance of the algorithm is evaluated by computing the percentages of sensitivity, specificity and accuracy. The data set is divided into two parts. In one part 15 patient records are considered as training set and the rest is to test the performance

SE (Sensitivity)= (TP / (TP+FN))*100

SP (Specificity)= (TN / (TN+FN))*100

AC (Accuracy)=(TP+TN) / (TN+TP+FN+FP)*100

Where TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Fig. 3 Complexity measures

TABLE IIII
ANALYSIS RESULT

Algorithm	Sensitivity	Specificity	Accuracy
ID3 Algorithm	92%	93.5%	94%

TABLE IVV
CLINICAL PROFILE OF 20 PATIENTS

	No of Cases
A. **Age**	
11-20	7
21-30	10
31-40	2
41-50	1
B. **Gender**	
Male	2
Female	18
C. Mucocutaneous Manifestation	
Photosensitivity	14
Malar rash	11
Alopecia	18
Oral Ulcers	17
Raynaud's symptom	9
Vasculitic rash	9
E. **Immunological profile**	
ANA	19
Anti-dsDNA	11
F. **Menstrual irregularity**	
Menarche not attained	10
G. **Survival**	
Regular follow up	17
Lost to follow up	4
Died	8
H. Haematological	
Anaemia	16
Leucopenia	7
Thrombocytopenia	10
I. **Musculoskeletal**	
Polyarthritis	11
Oligoarthritis	12
Monoarthritis	1
Myalgia	17

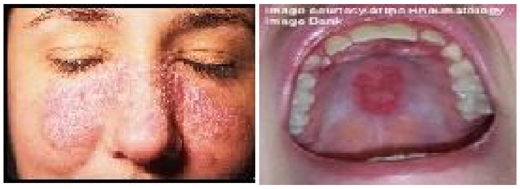

Fig. 4 Sample Lupus disease causes

Fig 4 shows the symptom of lupus disease. The left picture shows the malar rash and right side is oral ulcer symptom.

V. Conclusions

Health care and disease related data are voluminous and they are diverse in nature. The patients need special care and diagnosis in order to predict the disease earlier. Thus this paper will provide an efficient and new methodology to predict the chronic lupus disease and which in turn will extend the survival rate of the patients.

References

[1] A. T. Sayad, P.P. Halkarnikar, " Risk level prediction system in ischaemic heart disease", proceedings of IRF International conference, 30th March-2014, ISBN:978-93-82702-69-6

[2] S. Vijiyarani, S. Sudha, " Disease prediction in data mining technique – A Survey", International journal of computer applications and information technology, Vol 2, issue 1, Jan 2013

[3] Vikas Chaurasia, Saurabh pal, " Early prediction of heart disease using data mining techniques", Caribbean journal of science and technology, vol 1, 2013

[4] Renu saigal, Amit kansal, Manoop mittal, Yadvinder singh, Hari ram maharia, Manish Juneja, " Journal, Indian academy of clinical medicine, vol 13, no 1, March 2012.

[5] V. Manikandan, S. Latha, "predicting the analysis of heart disease symptoms using medical data mining methods", International journal on advanced computer theory and engineering, Vol 2, Issue 2, 2013.

[6] http://www.lupus.org

[7] http://www.medify.com

[8] http://pharmgkb.org

[9] http://www.ezdi.us

[10]
http://www.lupusresearchinstitute.org

[11] http://www.rheumatology.org