Volume 4, Issue 3, March 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Data Mining Classification Approach to Predict Systemic Lupus Erythematosus using ID3 Algorithm

Gomathi. S*
Information and Computer Technologies,
Sri Krishna Arts and Science College, India

Dr. V. Narayani
Associate Professor, Dept. of MCA
Karpagam College of Engineering, India

Abstract— Systemic Lupus Erythematosus is also known as SLE or lupus is an autoimmune disease which affect any part of the human body. The course of the study is unpredictable. Predicting SLE is a hectic task. Data mining classification approach will be better to predict the disease easily. This paper deals about the deadly disease SLE and a effective way to predict and analyse the disease. A new framework is proposed for the diagnosis and predicting the disease earlier will be used to extend the life of patient with lupus.

Keywords — Lupus, antinuclear antibody, oral ulcers, ID3, weka, American college of rheumatology.

I. INTRODUCTION

Health care field contain huge amount of data which holds sensitive information about patient details like diagnosis, prognosis and medical conditions. Data mining classification techniques can discover latent patterns or hidden relationships from the medical data sources. The medical environment is rich in information but poor in knowledge [1]. There are diseases on which people have no awareness about it. Prediction and analysis is difficult among practitioners and doctor. Some diseases are analysed at the end or in the extreme stage. One among that type of disease is Lupus also called as Systemic Lupus Erythematosus [13]. SLE is a complex autoimmune chronic disease which can be affected by environmental or genetic factors [11]. Lupus affects 10 women than men and the ration will be as 10:1. SLE causes inflammation, pain, swelling and rashes. It affects all parts of the body like kidney, lungs, heart, joints, nervous system and skin. Lupus condition varies from mild to serious. Lupus occurs 10 times more often in women and the disease is more common in some ethnic groups like blacks and Asians, tends to be worse in these groups. SLE is spreading in India too [12]. There is no cure for lupus and the treatment will be challenging. This disease can be predicted earlier using the classification algorithm. In this paper ID3 classification algorithm is used to predict the disease in early stage [2]. American lupus foundation has suggested eleven criteria to categorize and diagnose lupus.

II. LITERATURE REVIEW

HianChyeKoh and Gerald Tan discussed about data mining and its applications with major areas on treatment effectiveness, detection of fraud, Management of healthcare, detecting abuse, Customer relationship management in health care domain [5]. M. Durairaj, K. Meena describes a hybrid prediction system with Rough Set Theory and Artificial Neural Network used to predict and classify medical data. The new data mining technique and software to assist challenging solutions for medical data analysis has been explained. They proposed a hybrid tool that with RST and ANN to make effective data analysis and indicative predictions. The prediction accuracy and reliability is recorded by comparing observed and predicted rate[6].

Lupus is a Latin word which means ‘wolf’ which was first used during the middle ages which is to indicate skin lesions caused by a ‘wolf’s bite’ [8]. Ferdinand von Hebra in a Viennese physician in 1846 (1816–1880) introduced the butterfly metaphor to describe the malar/butterfly rash and he also used the term ‘lupus erythematosus’ and published his first illustrations in Atlas of Skin Diseases in 1856 [9]. Moriz Kaposi (1837–1902) was first recognised as a systemic disease with visceral manifestations called Lupus. Osler in Baltimore and Jadassohn in Vienna further established the systemic form. Reinhart and Hauck from Germany (1909) describe the other important milestones which includes the description of the false positive test for syphilis [10].

III. PROBLEM SPECIFICATION

There is no specific classification technique or prediction tool to predict SLE. People are not aware about the disease. SLE is slowly spreading in India where doctors and practioners don’t have effective technique to predict the disease in earlier stage. Thus in this paper ID3 algorithm is proposed to classify and predict the disease effectively. A framework is proposed to show functioning of the system.

IV. PROPOSED WORK

A. Eleven ACR conditions for SLE

American College of Rheumatology revised criteria for classification of systemic lupus erythematosus. The ACR has a list of symptoms and various measures that doctors can use as a guide to decide if a patient with symptoms has lupus. If

doctor finds that patient have at least four of the above mentioned problems, and finds no other reason for them, the patient may have lupus.

TABLE I
ELEVEN ACR CONDITIONS [8]

S. No	American College of Rheumatology	
	Symptoms	**Description**
1	Malar Rash	Fixed erythema, flat or raised, over the malar eminences, tending to spare the nasolabial folds
2	Discoid rash	Erythematous raised patches with adherent keratotic scaling and follicular plugging; atrophic scarring may occur in older lesions
3	Photosensitivity	Skin rash as a result of unusual reaction to sunlight, by patient history or physician observation
4	Oral ulcers	Oral or nasopharyngeal ulceration, usually painless, observed by physician
5	Nonerosive arthritis	Involving 2 or more peripheral joints, characterized by tenderness, swelling, or effusion
6	Pleuritis or pericarditis	Pleuritis--convincing history of pleuritic pain or rubbing heard by a physician or evidence of pleural effusion Or Pericarditis--documented by electrocardigram or rub or evidence of pericardial effusion
7	Renal disorder	Persistent proteinuria > 0.5 grams per day or > than 3+ if quantitation not performed or Cellular casts--may be red cell, hemoglobin, granular, tubular, or mixed
8	Neurologic disorder	Seizures--in the absence of offending drugs or known metabolic derangements; e.g., uremia, ketoacidosis, or electrolyte imbalance or Psychosis--in the absence of offending drugs or known metabolic derangements, e.g., uremia, ketoacidosis, or electrolyte imbalance
9	Hematologic disorder	Hemolytic anemia--with reticulocytosis or Leukopenia--< 4,000/mm^3 on ≥ 2 occasions or Lyphopenia--< 1,500/ mm^3 on ≥ 2 occasions or Thrombocytopenia--<100,000/ mm^3 in the absence of offending drugs
10	Immunologic disorder	Anti-DNA: antibody to native DNA in abnormal titer or Anti-Sm: presence of antibody to Sm nuclear antigen or Positive finding of antiphospholipid antibodies on or an abnormal serum level of IgG or IgM anticardiolipin antibodies, with positive test result for lupus anticoagulant using a standard method, or a false-positive test result for at least 6 months confirmed by Treponema pallidum immobilization or fluorescent treponemal antibody absorption test
11	Positive antinuclear antibody	An abnormal titer of antinuclear antibody by immunofluorescence or an equivalent assay at any point in time and in the absence of drugs

B. Images of some common SLE conditions

Fig. 1 Rashes on nails – Discoid rashes (criteria 2 by ACR) [7]

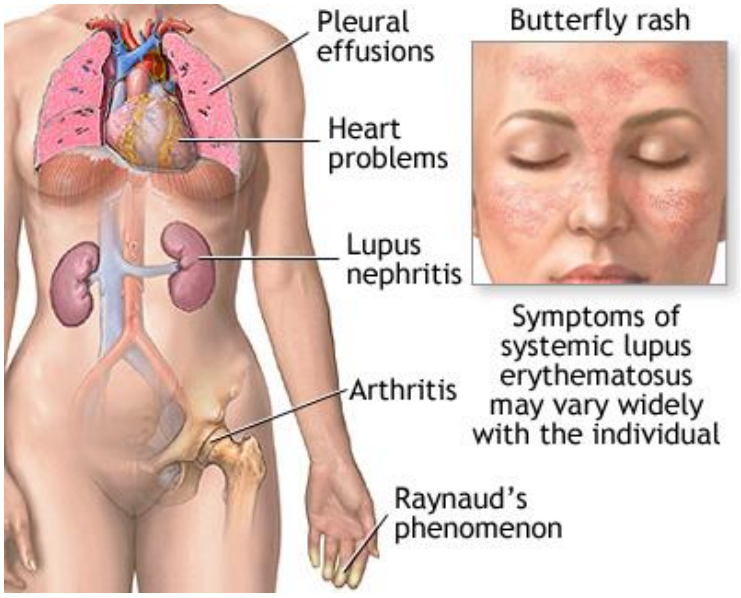

Fig. 2 Various parts of body where SLE will affect [7]

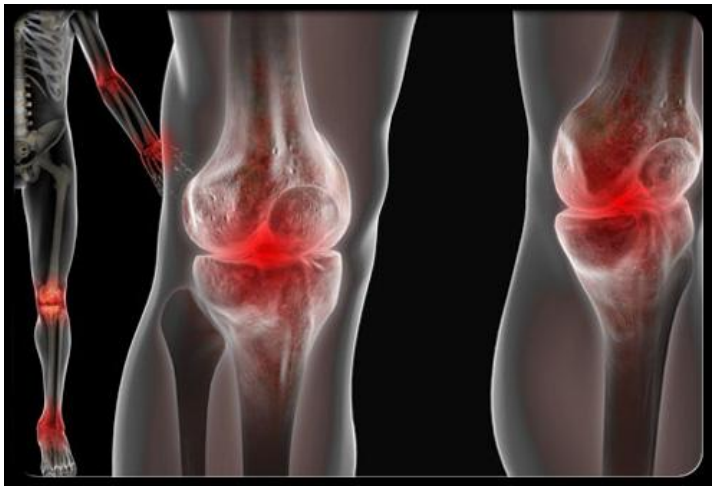

Fig. 3 Non erosive arthritis. Pains in joints.

Fig 1, 2, and 3 shows the sample images where SLE will affect the human body.

C. ID3 classification algorithm

ID3 stands for Iterative Dichotomiser 3 is used to generate a decision tree and predecessor of C4.5 Algorithm. The algorithm will work best on a set of training data, which is then sorted into classes, finally creates a tree which divides the data with common attributes. The tree is then used to classify real-world medical data of SLE patients of the same variety. If the four conditions of ACR are classified then the patient is said to be affected with lupus [8]. The advantages of ID3 algorithm are (i) easily implemented, (ii) a simple process, and (iii) running time increases only linearly with the complexity of the problem. Some function is necessary to measure what type of questions provides the most balanced splitting which is information gain metric [3].

D. Entropy

Entropy is an important concept to define information gain. Entropy is a measure of the impurity in a collection of training sets. A set S, containing these positive and negative targets, the entropy S related to the Boolean classification is:

$$E(S) = -P(pos)\log_2 P(pos) - P(neg)\log_2 P(neg)$$

Where E is entropy, P(pos) is proportion of positive samples in S and P(neg) is proportion of negative samples in S.

E. Information Gain

To minimize the depth of the decision tree while traversing the tree path, to select the optimal attribute that used to split the tree node, which can easily imply that the attribute with the most entropy reduction. Information gain is defined as the expected reduction of entropy related to specified attribute while splitting a decision tree.

$$Gain(S,A) = Entropy(S) - \sum_{v=1}^{n} (|Sv|/|S|) * Entropy(Sv)$$

The main thought about gain is to rank attributes and to build decision tree where at each node is located the attribute with greatest gain among the attributes not yet considered in the path from the root [3].

F. ID3 Pseudo code

ID3 (Learning Sets S, Attributes Sets A, Attributes values V) Return Decision Tree.
Begin
Load learning sets first, create decision tree root node 'rootNode', add learning set S into root node as its subset.
For rootNode, compute Entropy(rootNode.subset) first
If Entropy(rootNode.subset)==0, **then**
rootNode. subset consists of records all with the same value for the categorical attribute,
return a leaf node with decision attribute: attribute value;
If Entropy(rootNode.subset)!=0, **then**
Compute information gain for each attribute left (have not been used in splitting),
find attribute A with **Maximum(Gain(S,A))**.
Create child nodes of this rootNode and add to rootNode in the decision tree.
For each child of the rootNode, apply
ID3(S, A, V) recursively until reach node that has entropy=0 or reach leaf node.
End ID3.

G. Proposed framework

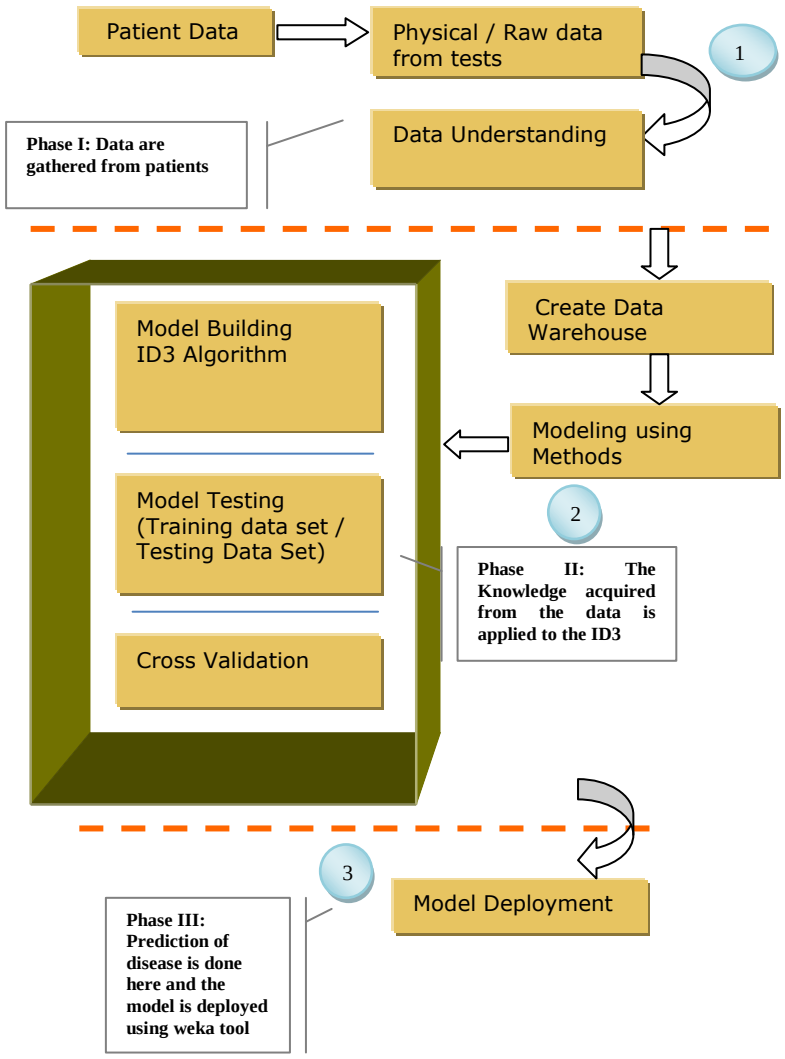

Fig. 4 Proposed Framework to predict SLE using ID3

A new framework for the prediction of SLE is proposed in Fig 4. The proposed work is categorized into three phases. Phase I is the process of gathering data. The patient's data is gathered and the tests undergone by the patient is analyzed well. The output of the phase I is given as the input to the Phase II. In Phase II the classification of the disease is made

using ID3 algorithm on the training set. In Phase II the data are tested carefully to check whether the prediction is accurate about the disease. The cross validation is done in the training set. Based on the accuracy the original data is tested and finally SLE is predicted and a model is deployed in Phase III.

H. Weka tool – Advantage of using this tool to mine data.

In Weka tool, routines which can be used stand alone via the command line are implemented as classes and arranged in packages. The tool comes with an extensive Graphical User Interface (GUI). It uses flat text files, work with a wide variety of data files including its own .arff and C4.5 file formats, data can be imported from a file in various formats like ARFF, CSV, C4.5, binary and data can also be read from a URL/ from an SQL database (using JDBC). Classifiers in WEKA are models for predicting nominal or numeric quantities [7].

V. CONCLUSION

The paper suggested an effective algorithm to diagnose SLE and a framework has been proposed. Yet no data mining classification approach has been proposed for the prediction of Lupus. Since this disease will pretend like other disease, current manual diagnosis fails to predict SLE in its starting stage. Since the prediction of disease is not so easy like other common disease like cancer, heart attack etc., a special classification technique is used and a modern framework has been proposed for the prediction purpose. The future work will be to apply this technique in a set of data set in weka data mining tool and to develop a model and finally to deploy the data sets to predict the disease.

REFERENCES

[1] Tom M. Mitchell, (1997). Machine Learning, Singapore, McGraw Hill.
[2] Asma Shaheen, Waqas Ahmad khan, "Intelligent Decision Support System in Diabetic eHealth Care from the perspective of Elders", Master Thesis Computer Science, Thesis no: MCS-2009-20, June 2009.
[3] Dipti D. Patil, V.M. Wadhai, A. Gokhale, "Evaluation of decision Tree Pruning Algorithms for Complexity and Classification Accurac", Int.ernational Journal of Advance Soft Computing Application, Vol. 2, No. 1, March 2010 ISSN 2074-8523, 2010
[4] Asha Rajkumar, G.Sophia Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm", Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Version 1.0 September 2010.
[5] HianChyeKoh and Gerald Tan, "Data Mining Applications in Healthcare", journal of Healthcare Information Management, Vol 19, No 2.
[6] M.Durairaj, K.Meena, "A Hybrid Prediction System Using Rough Sets and Artificial Neural Networks", International Journal Of Innovative Technology & Creative Engineering (ISSN: 2045-8711) Vol.1, No.7 July 2011.
[7] http://www.cs.waikato.ac.nz/ml/weka - Weka: Data Mining Software in Java,
[8] http://www.rheumatology.org/practice/clinical/classification/SLE/sle.asp
[9] http://aje.oxfordjournals.org/content/145/5/408.short
[10] http://journals.lww.com/md-journal/Abstract/2003/09000/Morbidity_and_Mortality_in_Systemic_Lupus.2.aspx
[11] http://www.healthline.com/health/systemic-lupus-erythematosus?toptoctest=expand
[12] http://www.nature.com/nrrheum/focus/lupus/index.html
[13] http://gerson-research.org/docs/HildenbrandGLG-1992-4/

