DATA MINING THE LUPUS DISEASE

Ms.S. Gomathi[1], Dr. V. Narayani[2]

[1]*Assistant Professor, Department of ICT, Sri Krishna Arts and Science College, Coimbatore, (India)*

[2]*Director I/C, Department of MCA, Karpagam College of Engineering, Coimbatore, (India)*

ABSTRACT

Lupus is chronic disease which affects almost all parts of the body. Predicting the disease is not as easy as it damages the tissue, organs and also imitate like other disease. Thus it needs a special algorithm to predict in earlier stage. Data mining is one of the best traditional methods to extract the hidden knowledge from large and voluminous data. Classification is one of the best techniques which is more often used to predict disease. Thus prediction can be done easier using data mining classification technique. Lupus is auto immune disease which cannot be cured. The disease can be predicted before it spread severely. There are various classification techniques in data mining. In this paper Classification and Regression Tree (CART), a decision tree is used to classify the data set and predicts the disease efficiently. This paper discusses Cartesian algorithm and attributes to analyze 45 lupus patients which are effectively applied to predict the lupus.

***Keywords*: CART, Data Mining, Disease, Auto Immune, Kidney Biopsy, Arthritis, Antigens, Autoantibody, Weka, Lupus.**

I. INTRODUCTION

Lupus is deadly disease which attacks the immune system and don't have any cure. The immune system is made up of white blood cells (White Blood Cells) which protect our body from foreign invaders (Virus, bacteria etc.,). The immune system generates antibodies to protect and destroy viruses, germs and bacteria. Lupus causes auto-anti bodies which can't identified by foreign invaders and healthy tissue which in turn causes the chemical reaction on the life time of the lupus patients can be extended once they are diagnosed. The studies show that the affect of lupus is more on females [6,23]. The major disorders of lupus are immunological disorder, hematological disorder, neurologic disorder, renal disorder and other symptoms are serositis, arthritis, oral ulcers, photo sensitivity and malar rash. The recent study shows that men will have less photosensitivity than women [7, 21]. Mining the data is a process of examining the large dataset which is routinely collected [15]. The technique is widely used in exploratory analysis [17]. Predicting the disease earlier has become a challenging task in medical field since medical data are voluminous and homogeneous in nature [2, 22]. Mining lupus is a challenging one which involves many attributes to diagnose the disease. Lupus cannot be diagnosed with a single lab test or image scanning methodology [8]. The combination of 2 or more test can conclude the occurrence of lupus disease. Mostly the disease is common in Hispanic, Asian, American and now it's common in India. There is no proper awareness about the disease. The causes of disease is not clear and which can be due

IJARSE

to gene, environment etc., [9] The lab test to diagnose lupus includes antinuclear antibody, kidney biopsy, c reactive protein, anti sm-double DNA etc.,

II. LITERATURE REVIEW

Jyothi et.al., [1] used three different supervised learning algorithms Naïve Bayes, decision tree and K-NN which is used to analyze the heart disease. The author used Tanagra tool to classify the data which is then evaluated using 10 fold cross validation.

Sandhiya et.al., [2] suggested Naïve Bayes which is used to predict attributes such as age, sex, sugar, blood pressure etc., She also suggested CART algorithm. The author analyzed 500 patients' data sets. The data was analyzed using WEKA tool and the results are tabulated.

Manikandan et al., [3] used ANN, decision tree and logistic regression techniques for heart prediction analysis. The author used data of 20 variables in the prediction model. Authors compared and found the efficiency and accuracy of decision tree is higher than ANN.

Shira et al., [4] presented that the main involvement of lupus will be on brain (Central Nervous System). The modalities are classified into functional and morphological. The image results can be acquired from Magnetic Resonance Imaging (MRI), Diffusion Weighted Imaging (DWI), Diffuse Tensor Imaging (DTI) and Magnetic Resonance Spectroscopy (MRS).

John et al., [5] outlines the patterns of lupus in particular subset of patients based on age, ethnicity, gender and social class. The author classified the patients based on the severity of the organs involved and the tests undergone.

III. PROBLEM SPECIFICATION

The awareness about the disease severity and occurrence of lupus is not much popular among most of the people. Lupus affects more women than men. American College of Rheumatology summarized 11 criteria to analyze the existence of lupus. Many hospitals lack in predicting the disease earlier. Since the disease act like many diseases, prediction is so challenging in medicinal field. It affects any part in any sequence like skin, lungs, kidney, heart etc., An effective algorithm CART is used in this paper to predict the disease effectively. CART is a decision tree algorithm; a decision tree is generated to show how the analysis and prediction is done. Important attributes to predict lupus is tabulated and 45 patients have been analyzed and diagnosed using CART.

IV.DATA ANALYSIS AND PREPARATION

The most important methodology implemented in this paper is CART algorithm. The data study consists of lupus dataset which includes list of attributes and the explanation of the attributes is tabulated in Table I. The database is collected from a biological lab. 45 patients have been analyzed based on the gender, age, ethnicity, history of patients detail etc.,. The result is shown and the patients are analyzed and categorized in Table III. The data is interpreted using weka tool and the results are depicted in Fig 5 to Fig 10.

IJARSE

V.A GLANCE ABOUT LUPUS

Fig. 1: Neonatal Lupus which affect the child. Rashes on the skin show the presence of lupus. But people normally consult the dermatologist to cure the rashes [13,20].

Fig 2: The tared and reddish nail shows the symptom of lupus. This may also happen if there are some skin problem. But this may also a symptom for lupus [10,11].

Fig 3: Subacute lupus: The lupus which affect the skin. The red patches in the hand shows the symptom of lupus [12].

VI.ATTRIBUTES TO DIAGNOSE LUPUS

TABLE I shows the Selected attributes which is used to diagnose the patients. This is analysed from the huge records and case sheet of patients with lupus. Many attributes and tests are involved in predicting the lupus. The below attributes are important to predict the disease[14,16,23,24]. The ACR criteria is given as the 5th attribute which must be analysed during the treatment of the patient.

Table I. Attributes To Diagnose Lupus

Predictable Attribute	
Diagnosis (value 0: satisfies any 4 ACR criteria (have lupus but chance of extending the life); value 1: satisfies more than ACR 4 criteria (have lupus difficult to extend the life); value 2: not satisfied any 4 criteria (less possibility of lupus, may have any other problem, consult rheumatologist)	
Key Attribute	
Patient ID – Patient;s identification number	
ID	**Input Attributes**
1.	Age (Value 0: >35; value 1: <=35 && > 20; value 2: <=20)
2.	Sex (value 0: male; value 1: female)
3.	Sample Type (value 0: Serum; value 1: Plasma; value 2: urine)
4.	Ethnicity (value 0: African; value 1: American; value 2: Hispanic; value 3: Asian; value 4: Caucasian)
5.	ACR Criteria (value 1: malar rash; value 2: discoid rash; value 3: photosensitivity; value 4: oral ulcers; value 5: non erosive arthritis; value 6: pleuritis; value 7: renal disorders; value 8: neurologic disorder; value 9: hematologic disorder; value 10: immunologic disorders; value 11: antinuclear antibody
6.	Disease activity (value 0: mild; value 1: moderate; value 2: severe)
7.	Tests involved (value 0: ANA; value 1: CBC; value 2: Chest X-ray; value 3: Kidney biopsy; value 4: Urinalysis; value 5: Rheumatoid test facts; value 6 :Liver function blood test; value 7: ESR)

VII.CART

Classification tree is used for categorical/nominal target variable. Gini index is the impurity measure which is used to select the variable in CART [18].

Impurity Measures In CART

There are four impurity measures (i), GINI, (ii) Twoing, (iii) Least Square Deviation and (iv) Ordered towing [19]. The GINI and twoing is used for categorical variables and least squared deviation is used for continuous target variables. Thus to predict this disease GINI and twoing is used.

GINI Index

This index is used to measure the inequality and uneven distribution of the variables. The range of gini index is between 0 to 1, where 0 implies perfect equality and 1 implies perfect inequality [20].

$$GINI\ (t) = \sum_{j<>1} p(j/t)p(i/t) \qquad (1)$$

Where t is the root node, i,j are the categories of target variable and p(j/t) is the proportion of the target category j in node t.

IJARSE

Twoing

Twoing is splitting target variables into two super classes and finding the best split on the predicted variable based on super class [22].

$$GINI_{Twoing}(s,t) = P_L.P_R/4\ [\ \sum_j\ P(j/t_L) - P(j/t_R)\ ]^2 \qquad (2)$$

Where s – split, t – node, t_L and t_R – node created by split.

VIII.CART ALGORITHM

Input: t,s

Output: Perfect tree to predict

Variables: t – root node, s- variables

begin

Step 1: start from t=1
Step 2: Search for split with set of variables
Step 3: Split the node into two nodes.
Step 4: Repeat split searching process
Step 5: Continue tree growing

end

IX.EXPERIMENTAL RESULT

The performance of the algorithm is evaluated with three other classification algorithms (Decision tree, Support Vector Machine and ID3) by computing the percentages of sensitivity, specificity and accuracy. The data set is divided into two parts.

Table II. Performance Evaluation Of Cart With Other Algorithms

Classification Technique	Specificity	Sensitivity	Accuracy
Decision Tree	93%	93.5%	94%
ID3	92%	93.2%	93.5%
CART	93.5%	94%	94.5%
SVM	90%	89%	91%

The reason for why CART is implemented to predict lupus is shown in the table. CART has the highest specificity, sensitivity and accuracy to yield the result compared to other classification techniques.

X.TREE STRUCTURE OF CART ALGIRTHM

Fig.4. Shows the Anti nuclear antibody (ANA) is the test which is important to be taken to find the occurrence of lupus. 90% of positive ANA may yield to the presence of lupus. If it is positive means, the ACR criteria will be diagnosed further with the other lab tests and prognosis. If it is negative means the patients may suffer with some other disease and he/she will be suggested to the rheumatologist.

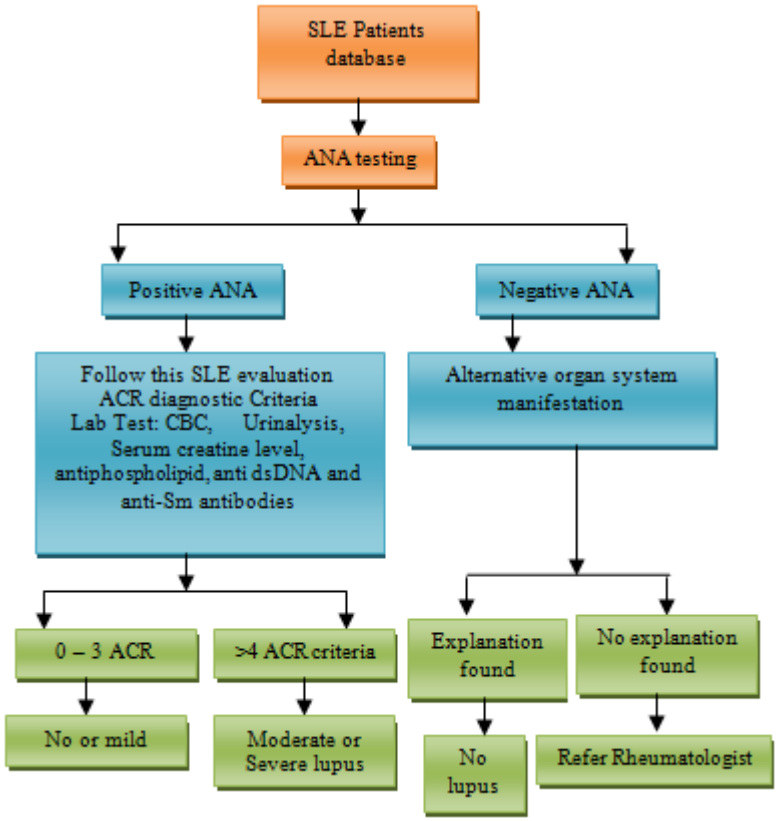


Fig. 4: CART tree generated to predict lupus

IJARSE

Table III. Profile Of 45 Patients

	No of Cases
Age	
>35	30
<=35 && >20	10
<=20	5
Gender	
Male	3
Female	42
Mucocutaneous Manifestation	
Photosensitivity	20
Malar rash	15
Alopecia	16
Oral Ulcers	14
Raynaud's symptom	10
Vasculitic rash	10
Immunological profile	
ANA	41
Anti-dsDNA	35
Survival	
Regular follow up	25
Lost to follow up	12
Died	8
Haematological	
Anaemia	15
Leucopenia	10
Thrombocytopenia	13
Musculoskeletal	
Polyarthritis	12
Oligoarthritis	10
Monoarthritis	12
Myalgia	20

TABLE III shows the clinical profile of 45 patients. The patients who have the symptoms such as mucocutaneous (skin portion), immunological affect, causes in blood, bones etc are shown with the count.

XI. RESULTS: CLINICAL PROFILE OF 45 PATIENTS

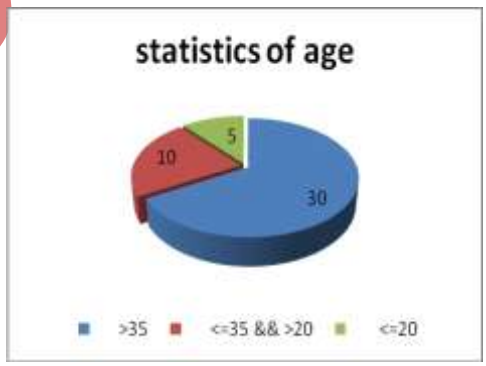
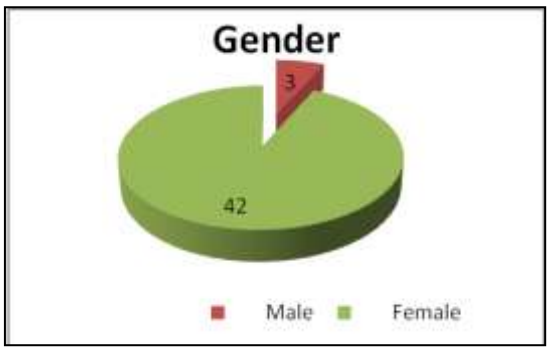


Fig.5: Shows the causes of lupus based on age.

IJARSE

From Fig 5 it is clear that 45 patients has been analyzed, most of the lupus symptoms is diagnosed above 35 years of age. Other ages are between 20 to 35.



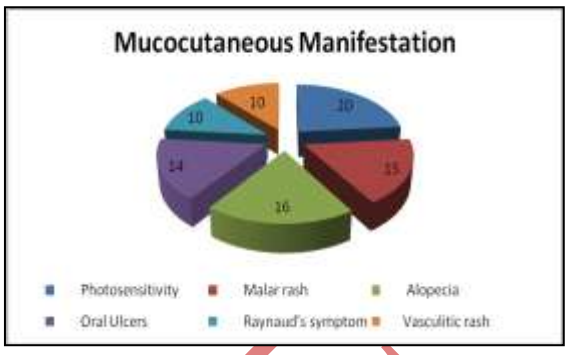
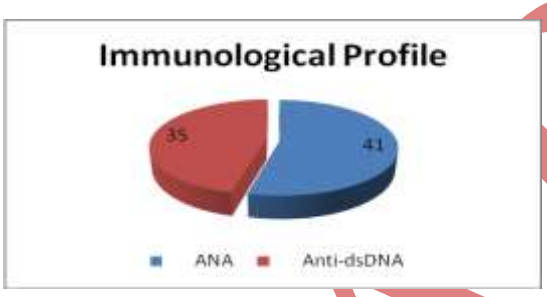


Fig. 6: Shows the causes of lupus based on gender **Fig 7: Causes of lupus in Skin**

Fig 6 shows that most, female suffer with lupus more than men. 42 females and 3 males suffered with lupus. The reason why women suffer more than men is, the lupus will aggravate more during menstrual period.

From fig 7, it is clear that most of the lupus patients suffer with photosensitivity (i.e.,) affect of sun and the other symptoms are alopecia, malar rash, oral ulcers, raynaud's symptoms (blue colored finger) and vasculitic rash.



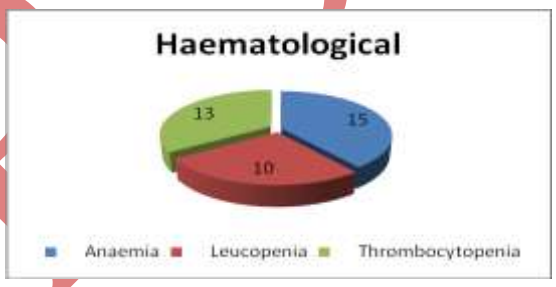


Fig. 8: Lab tests (immunological test) **Fig 9: Symptoms on Blood (haematological disorder)**

The two main test to be taken to predict lupus is ANA (Anti nuclear antibody) and Anti-dsDNA. If the person is tested with positive ANA then there is 95% chance of lupus. From the fig 8, it is clear that 41 patients have positive ANA and 35 patients have Anti-dsDNA.

Anaemia, leucopenia and thrombocytopenia are the main Symptoms of lupus which affect blood. Lupus disease will affet tissue, blood even bones.

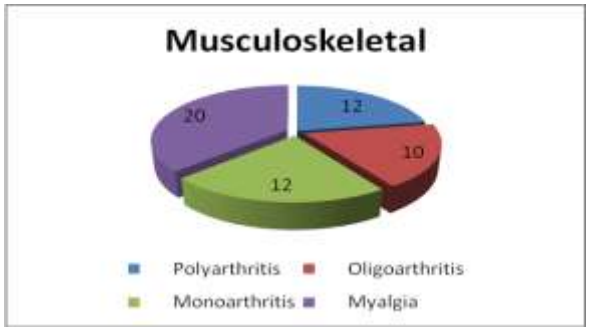


Fig.10: Symptoms of lupus on bones (musculoskeletal)

IJARSE

Fig 10 shows that 20 patients affected with polyarthritis, 12 affected with myalgia and monoarthritis and 10 patients affected with oligoarthritis. This shows the presence of lupus on bones.

XII. CONCLUSION

Though the medical data is huge and voluminous, there are some techniques which can be integrated to extract the data to predict the disease. Lupus can be diagnosed and predicted which cannot be completely cured. The prediction should be done earlier to extend the life of the patient. Thus it needs a special technique to predict this kind of autoimmune disease. The paper outlined about how to predict the lupus disease with its diagnosing criteria, proposed attributes and algorithm. CART algorithm is implemented in to diagnose lupus accurately and efficiently. A decision tree is generated using CART algorithm. Various important measuring criteria like GINI index and twoing index to measure CART is given.

REFERENCES

[1] Jyoti Soni, Ujma Ansari, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Predicition," International Journal of Computer Applications, Vol 17, No-8, March 2011.

[2] Sandhiya, Pavithra, et.al.,"Novel approach for Heart Disease Verdict Using Data Mining Technique", International Journal of Modern Engineering Research.

[3] Manikandan. V, Latha. S, "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.

[4] Shira, Ilan Asher et al., "Novel Biological Treatments for Systemic Lupus Erythematosus: Current and Future Modalities," IMAJ Reviews, August 2012.

[5] John A Reynolds, Ian N Bruce, "Overview of the management of Systemic Lupus Erythematosus", Arthritis Research UK Topical Reviews, Spring 2013.

[6] http://en.wikipedia.org/wiki/Lupus_erythematosus.

[7] http://www.medicalnewstoday.com/info/lupus/

[8] http://www.patient.co.uk/health/systemic-lupus-erythematosus-leaflet

[9] https://www.womenshealth.gov/publications/our-publications/fact-sheet/lupus.html

[10] http://www.lupusresearch.org/public-policy/resources/faqs_lupus.html

[11] http://lupusimages.com/

[12] http://www.rheumatology.org/practice/clinical/classification/SLE/sle.asp

[13] A. Kumar, Indian Guidelines on the management of SLE, Journal of Indian Rheumatol Association 2002.

[14] Tsokos GC. Systemic lupus erythematosus . N Engl J Med 2011; 365: 2110-21.

[15] Arvind Sharma and P.C. Gupta , Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool International Journal of Communication and Computer Technologies Volume 01, No.6, Issue: 02 September 2012.

[16] Looney RJ, Anolik JH, Campbell D, et al. B cell depletion as a novel treatment for systemic lupus erythematosus: a phase I/II dose-escalation trial of rituximab.*Arthritis Rheum* 2004; 50 (8): 2580-9. Hingoran NG. High power electronics and flexible AC transmission system. IEEE Power Engineering Review July 1988:3–4.

IJARSE

[17] A. H. Mohamed and M. H. S. Bin Jahabar, Implementation and Comparison of Inductive Learning Algorithms on Timetabling, International. Journal of Information Technology., vol. 12, no.7, pp. 97–113, 2006.

[18] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Technique", 978-1-4244-1968-5/08/$25.00 ©2008 IEEE.

[19] K.P. Soman, Shyam Diwakar and V. Ajay, Insight into data mining theory and practice, Eastern economy edition, 2012.

[20] Renu Saigal, Amit Kansal, Manop Mittal, et.al., Clinical profile of Systemic lupus erythematosus patients at a teritiarycare centre in western India, JIACM, 2011.

[21] George Bertsias, Ricard Cervera, Dimitrios, Boumpas: Systemic Lupus Erythematosus: Pathogenesis and Clinical features. EULAR textbook on Rheumatic Disease.

[22] Srinivas: Novel approach for heart prediction verdict using data m ining technique. International Journal of Computer Science and Engineering. 2010.

[23] Gomathi.S, Dr. V. Narayani, "A Data Mining Classification Approach to Predict Systemic Lupus Erythematosus using ID3 Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 4, March 2014, ISSN: 2277 128X

[24] Gomathi.S, Dr. V. Narayani, "A predictive model for systemic lupus erythematosus disease using data mining", International Journal of Emerging Technologies and Applications in Engineering, Technology and sciences (ij-eta-ets) Vol 7, Jan-Jun 14, ISSN: 0974-3588

IJARSE