A Hybrid Clustering Data Mining Technique (HCDMT) for Predicting SLE

A. Malarvizhi, S. Ravichandran

Abstract*: *SLE is an auto immune and complex disease. Predicting Systemic Lupus Erythematosus (SLE) is significantly challenging due to its high level of heterogeneity in symptoms. There is a limitation on the tools used for predicting SLE accurately. This paper proposes a machine learning approach to predict the disease from SLE data set and classify patients in whom the disease is active. The data purified and selected for classification improves the accuracy of the proposed method called HCDMT (Hybrid Clustering Data Mining Technique), an amalgamation of CART and k-Means, was evaluated on SLE data. It was found to predict above 95% of SLE cases.

***Keywords* : *SLE, Clustering, Mining, Machine Learning*.**

.

I. INTRODUCTION

SLE has no definite diagnostic tools with which it can be determined due to its varying symptoms. Diagnostic approaches to SLE prediction have not changed for a long time making it a major therapeutic challenge. Clinicians rely on medical evaluations and specific laboratory tests which checks body complement levels or measures antibodies. Defects in body tolerance levels allow for activation of self-reactive B cell clones in SLE and generate plasma blasts that secrete auto-antibodies, damaging body tissues [1][2]. GWAS (Genome wide association studies) have identified many regulatory regions that influence these B cell function [3]. Auto antibodies with stimulate the production of interferon, an indication of infection in SLE all patients [4] [5]. Further, Myeloid cells (MC) also play a role in SLE pathogenesis [6]. Granulocytes with a low-density that are similar in structure to abnormal neutrophil cells, start appearing in SLE patient's blood

[7][8][9]. Though these cell have been linked in medical studies to vascular, kidney and other complications in SLE affected human body, they have not been extensively studied [10][11][12][13][14][15]. SLE disease's unpredictable activity has stalled from linking human body complications to the disease. Hence, improvements Data Mining techniques could be used for effective detection and treatment of SLE patients. Machine learning in Data mining has a wide range of computational methods which can help connect SLE's complex medical data to predictions of SLE and its state like active or inactive. Data mining techniques can be used to identify new biomarkers for SLE from urine or other tests used for identifying SLE [16][17]. SLE symptoms when studied over a period of time from tests can be used by machine learning algorithms to throughput common or similar patters exhibited by the disease and thus identify subjects that indicate higher degrees of SLE activity. This can also be a starting point for getting deep insights into SLE pathogenesis. Figure 1 depicts incidences of SLE

Fig. 1 – SLE (Rashes and Oral Cancer)

II. LITERATURE REVIEW

Machine learning in computer science is an emerging field that seeks to solve complex problems. SLE with its complexity due to multiple severity levels and symptoms is well suited to learning through machine learning architectures. Inductive learning algorithm (Decision Tree) implementing an intelligent system constructed decision trees using in java language for appropriate classifications. The study in [18] analyzed large volumes of health care data using rule based (C4.5), Decision tree and Naïve bayes classification methods for effective detection of heart attacks. The key events of pathogenesis of SLE were analyzed and environmental /hormonal factors in the disease were found and a classification criteria of lupus based on the limitations of diagnostics was proposed [19]. Many studies have addressed SLE by using clustering techniques to identify similar symptoms responsible for SLE [20] and its risk factors [21] where DNA of SLE patients collected in the laboratory was used for analysis. Though medical data offers huge amount data for study, very few cases of SLE data exist, making it difficult for an effective SLE predictive tool. Hence, this

Revised Manuscript Received on September 15, 2019.

* Correspondence Author

Ms. A. Malarvizhi, Research Scholar and Assistant Professor, PG and Research Department of Computer Science, H.H. The Rajah's College (A), Pudukkottai, Tamilnadu- 622001Email:malarselvamtnau@gmail.com

Mr. S. Ravichandran, Assistant Professor and Head, PG and Research Department of Computer Science, H.H. The Rajah's College (A), Pudukkottai, Tamilnadu- 622001 Email:rajahsravis@gmail.com

research work attempts to fill the gap by proposing a Hybrid Clustering Data Mining Technique called (HCDMT), an amalgamation of CART and k-Means for early SLE predictions.

III. HYBRID CLUSTERING DATA MINING TECHNIQUE CALLED (HCDMT)

Lupus in latin means 'wolf' and was initially used to specify erosive skin lesions caused by 'wolf's bite'. SLE affects almost all parts of the body like joints, skin, kidneys, nervous system, lungs ..etc. Lupus can be diagnosed with a minimum of two laboratory tests and clinical criteria. Tests can include antibody counts, blood count, urinalysis, Creatin levels and anti-double stranded DNA antibody. In the prevalence of dynamic and variations in data Data mining classification techniques are an efficient way to predict data. The proposed techniques is a combination of K-Means Clustering and CART. Decision Trees in data mining create a model that predicts a targeted value based on multiple inputs. CART was introduced by in 1984 [22][23]. Classification trees are used when the target variable is categorical and the tree identifies the class to which the target variable fits. Regression is used when it is continuous and tree is used to predict it's class value. CART algorithm is a sequence of questions and its answers determine the next question which results in a tree structure and ends in terminal nodes when there are no more questions. Figure 2 depicts a CART.

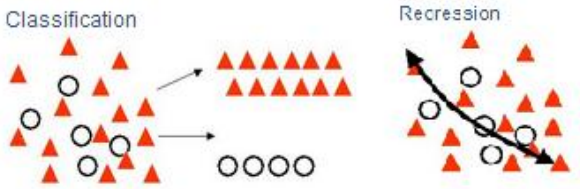

Fig. 2 - CART

K-means clustering is used on unlabeled data to categorize them and is an unsupervised learning technique. The data groups are found by a variable *K*. It is an iterative process where each point is assigned to cluster or group based on feature similarity. This results in creating clusters with centroids and labels for the unlabelled data. It uses distance as a metric to identify the distance of a data point from its cluster centroid. Figure 2 depicts K-Means Clustering

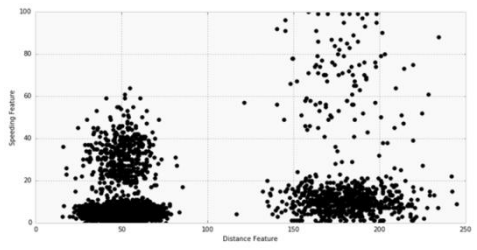

Fig. 3 – K-Means Clustering

CART is used for its simplicity to predict in large amount of data. Once required variables are predicted k-Means is used to identify clusters from the unlabelled data and thus effectively arrive at a conclusion of the disease. Figure 4 depicts the steps followed in the proposed **HCDMT.**

1. **Patients SLE raw data (Chicago Lupus Database (CLD))**
2. **Data Cleaning**
   a. **Missing value prediction**
   b. **Redundancy avoidance**
   c. **Filtering (Fill mean mode value)**
   d. **Attribute reduction**
3. **Decision Making (based on the condition to form a tree)**
   a. **Using decision tree algorithm**
4. **Ranking using Logical operators**
   a. **AND OR XOR conditions**
5. **Clustering based on Ranking method**
   a. **Hybrid algorithm using CART and K-MEANS**
   b. **In CART (i), GINI, (ii) Twoing, (iii) Least Square Deviation and (iv) Ordered towing**
6. **Disease Prediction**

Fig. 4- HCDMT Steps

The Dataset used is the study is the Chicago Lupus Database (CLD). This is a registry of individuals with lupus used for lupus research with a probable or definite lupus symptoms. .CLD attributes are Age, Gender, Test Sample, Disease Activity, Symptoms, Severity, Involved Organs, Tests conducted and follow-ups. Figure 5 depicts a screen shot of the data values.

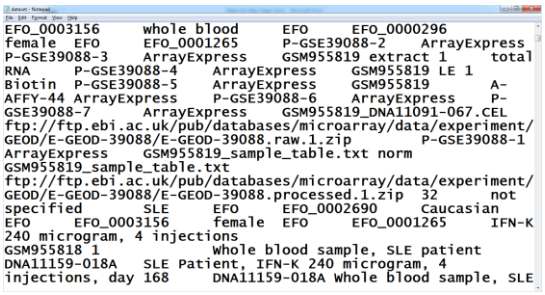

Fig. 5 – CLD Data Values

Real-world datasets can have missing values for various reasons and are often found as blanks or Not numbers or other placeholders. Such imperfect data can significantly impact a model's quality in terms of prediction output. Hence, HCDMT pre-processes data values of CLD with data cleaning by replacing missing values with averages based on similar kind of data records. Figure 6 depicts the screen shot of missing values

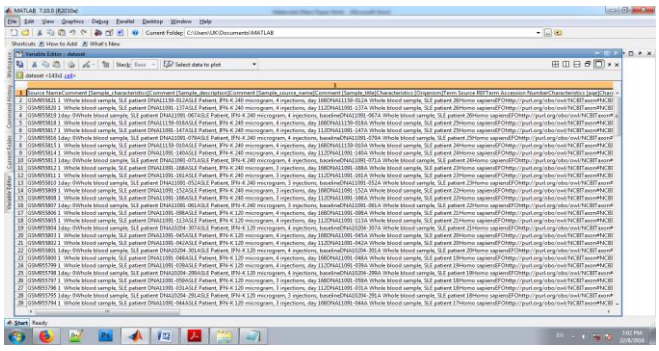

Fig 6 – Missing Values Prediction

Once missing values are replaced HCDMT follows it with removal of redundant data and filtering of unwanted values. Filtering involves Fill mean mode value when If 1 or 2 attribute values are missing and the mean mode value is filled in that corresponding row values. In Attribute reduction, unrelated attributes like PatientID, URL, ArrayIndex, Sample are removed. HCDMT uses several questions before using CART as listed in Table 1.

TABLE I: CONDITIONS USED BY HCDMT FOR CART

Attribute	Condition	Decision Value
Age	<=30	0
	>=31 and <=60	1
	>61	2
Gender	Female	0
	Male	1
Test Sample	Blood	0
	Plasma	1
	Urine	2
Disease Activity	Mild	0
	Moderate	1
	Severe	2
Symptoms	None	0
	malar rash	1
	discoid rash	2
	photosensitivity	3
	oral ulcers	4
	non erosive arthritis	5
	pleuritis	6
	renal disorders	7
	neurologic disorder	8
	hematologic disorder	9
	immunologic disorders	10
	antinuclear antibody	11
Severity	Low	0
	Medium	1
	High	2
Involved Organ	none	0
	Skin	1
	Joints	2
	Musculoskeletal	3
	Blood	4
	Brain	5
	Lung	6
	Central Nervous System	7
	Vascular	8
	Eyes	9
	Heart	10
	Pulmunory	11
	Gastrointensional	12
	Mouth	13
	extremities	14
Tests Taken	None	0
	AntiNuclear Antibody	1
	Complete Blood Count	2
	Chest X-ray	3
	Kidney biopsy	4
	Urinalysis	5
	Rheumatoid test facts	6
	Liver function blood test	7
	Erythrocyte Sedimentation Rate	8
Follow up	Regular	0
	Occasional	1
	None	2

The output of CART is then ranked by HCDMT. Ranking is based on higher combined values of the attributes, which can help forecast pre-dominant symptoms of SLE in CLD values. Figure 7 depicts a screen shot of HCDMT ranking.

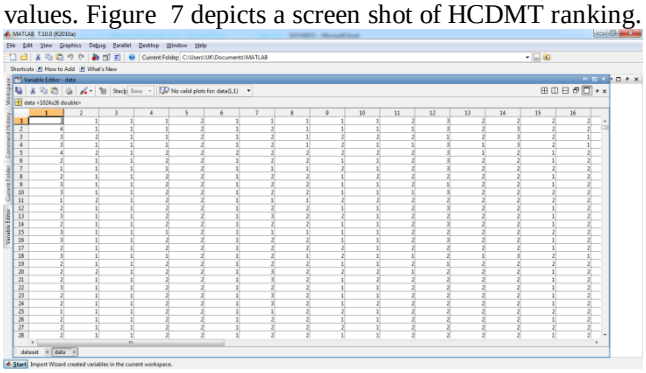

Fig. 7 – HCDMT Ranking

The out of CART and ranking is then clustered using K-Means in HCDMT to predict SLE. MATLAB was used in the entire process to achieve end results of HCDMT. Further, alternative DM techniques were also used to compare the proposed method in the same CLD database. Parameters of sensitivity, specificity and accuracy were used for the comparison between techniques. Table 2 lists comparative performances of techniques in predicting SLE.

TABLEII: COMPARATIVE PERFORMANCES OF SLE BY DM TECHNIQUES

Method	Sensitivity (%)	Specificity (%)	Accuracy (%)
CART	97.33	97.66	97.19
K-Means	93.2	93.1	93.5
Decision tree	96.1	97.3	96.8
Back propagation	97.13	97.86	97.45
HCDMT	98.33	98.66	98.45

To Validate the proposed HCDMT, it is evident from Table 2 that HCDMT performs better than CART (97.19), K-Means (93.5), Decision Trees (96.8), Back propagation of ANN (97.45) by scoring 98.45 in terms of predictive accuracy of SLE from CLD.

IV. CONCLUSION

SLE, a chronic autoimmune disease causes the human immune system to attack the body. Humans having SLE demonstrate several severity levels due to the complex interactions the disease triggers. This paper has proposed a novel hybrid technique using machine learning, a model capable of predicting SLE based on symptom with better accuracy. The most valuable outcome of the proposed model HCDMT is the actionable information can provide to physicians and patients. It can be concluded that HCDMT can be used as a promising and implementable technique for symptom management and SLE prediction in computer aided systems.

REFERENCES

1. Lugar, P. L., Love, C., Grammer, A. C., Dave, S. S. & Lipsky, P. E. Molecular characterization of circulating plasma cells in patients with systemic lupus erythematosus. PLoS One 7, e44362, https://doi.org/10.1371/journal.pone.0044362 (2012)
2. Karrar, S. & Cunninghame Graham, D. S. Abnormal B-cell development in systemic lupus erythematosus: what the genetics tell us. Arthritis Rheumatol. 70, 496–507 (2018).

3. Vaughn, S. E. et al. Lupus risk variants in the PXK locus alter B-cell receptor internalization. Front. Genet. 5, 450, https://doi.org/10.3389/fgene.2014.00450 (2015).
4. Bengtsson, A. A. & Rönnblom, L. Role of interferons in SLE. Best Pract. Res. Clin. Rheumatol. 31, 415–428 (2017).
5. Catalina, M. D., Bachali, P., Geraci, N. S., Grammer, A. C. & Lipsky, P. E. Gene expression analysis delineates the potential roles of multiple interferons in systemic lupus erythematosus. Communications Biology 2(1) (2019).
6. Labonte, A. C. et al. Identification of alterations in macrophage activation associated with disease activity in systemic lupus erythematosus. PLOS ONE 13(12), e0208132 (2018).
7. Hacbarth, E. & Kajdacsy-Balla, A. Low density neutrophils in patients with systemic lupus erythematosus, rheumatoid arthritis, and acute rheumatic fever. Arhritis Rheum. 29, 1334–1342 (1986).
8. Wright, H. L., Makki, F. A., Moots, R. J. & Edwards, S. W. Low-density granulocytes: functionally distinct, immature neutrophils in rheumatoid arthritis with altered properties and defective TNF signaling. J. Leukoc. Biol. 101, 599–611 (2017).
9. Scapini, P., Marini, O., Tecchio, C. & Cassatella, M. A. Human neutrophils in the saga of cellular heterogeneity: insights and open questions. Immunol. Rev. 273, 48–60 (2016).
10. Kegerreis, B. J. et al. Genomic Identification of Low-Density Granulocytes and Analysis of Their Role in the Pathogenesis of Systemic Lupus Erythematosus. The Journal of Immunology 202(11), 3309–3317 (2019).
11. Villanueva, E. et al. Netting neutrophils induced endothelial damage, infiltrate tissues, and expose immunostimulatory molecules in systemic lupus erythematosus. J. Immunol. 187, 538–552 (2011).
12. Lood, C. et al. Neutrophil extracellular traps enriched in oxidized mitochondrial DNA are interferogenic and contribute to lupus-like disease. Nat. Med. 22, 146–153 (2016).
13. Denny, M. F. et al. A distinct subset of proinflammatory neutrophils isolated from patients with systemic lupus erythematosus induces vascular damage and synthesizes type I IFNs. J. Immunol. 184, 3284–3297 (2010).
14. Jourde-Chiche, N. et al. Modular transcriptional repertoire analyses identify a blood neutrophil signature as a candidate biomarker for lupus nephritis. Rheumatology (Oxford) 56, 477–487 (2017).
15. Carlucci, P. M. et al. Neutrophil subsets and their gene signature associate with vascular inflammation and coronary atherosclerosis in lupus. JCI Insight 3, e99276, https://doi.org/10.1172/jci.insight.99276 (2018).
16. Wolf, B. J. et al. Development of biomarker models to predict outcomes in lupus nephritis. Arthritis Rheum. 68, 1955–1963 (2016).
17. Almlöf, J. C. et al. Novel risk genes for systemic lupus erythematosus predicted by random forest classification. Sci. Rep. 7, 6236, https://doi.org/10.1038/s41598-017-06516-1 (2017).
18. Srinivas, "Novel approach for heart prediction verdict using data mining technique", International Journal of Computer Science and Engineering. 2010.
19. Vikas chaurasia, Saurab pal, "Early prediction of heart diseases using data mining techniques", 2013 Vol 1, No 0799-3757
20. Armañanzas, R., Calvo, B., Iñaki, I., López-Hoyos, M., Martínez-Taboada, V., Ucar, E., …& Zubiaga, A. (2009, May 9). Microarray analysis of autoimmune diseases by machine learning procedures. IEEE Transactions on Information Technology in Biomedicine, 13(3), 341-350.
21. Ravenell, R., Kamen, D., Fleury, T. J., Spence, D., Hollis, B. W., Janech, M. G., …& Almeida, J. S. (2012). Premature Atherosclerosis Is Associated With Hypovitaminosis D and Angiotensin-Converting Enzyme Inhibitor Non-use in Lupus Patients. The American Journal of the Medical Sciences, 268-273.
22. http://statweb.stanford.edu/~olshen/
23. Manimaran R. and Vanitha M, "An Efficient Study on Usage of Data Mining Techniques for Predicting Diabetes", *International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)* Vol.3 (20), pp.268-272 ISSN: 2394-3785, 2016.