International Journal of Engineering & Technology, 7 (1.3) (2018) 28-32

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Early prediction of systemic lupus erythematosus using hybrid K-Means J48 decision tree algorithm

S. Gomathi [1]*, V. Narayani [2]

[1] Research Scholar, Department of Computer Science, Research and Development Centre, Bharathiar University, Coimbatore, India
[2] Director, MCA Department, Karpagam College of Engineering, Coimbatore, India
**Corresponding author E-mail: mailtogomathisrinivasan@gmail.com*

Abstract

The objective of the paper is to propose an enhanced algorithm for the prediction of chronic, autoimmune disease called Systemic Lupus Erythematosus (SLE). The Hybrid K-means J48 Decision Tree algorithm (HKMJDT) has been proposed for the effective and early prediction of the SLE. The reason for combining both the clustering and classification algorithms is to obtain the best accuracy and to predict the disease in the early stage. The performance of algorithms such as Naïve Bayes, decision tree, random forest, J48 and Hoeffding tree has been combined with K-means clustering algorithm and compared in an effort to find the best algorithm for diagnosing SLE disease. The results of the statistical evaluation with the comparative study show that the effectiveness of different classification techniques depends on the nature and intricacy of the dataset used. K-means combined with J48 algorithm shows the best accuracy rate of 82.14% on the pre-processed data. The work-flow has been proposed to show the execution of the algorithm.

***Keywords*:** *Use Data Mining; Auto-Immune; Lupus; J48; K-Means; Classification; Clustering; Chronic; Decision Tree; Sensitivity; Specificity; Accuracy.*

1. Introduction

The article in 1983 briefed the concept of SLE with the clinical laboratory features, disorders, treatments, causes and pathophysiology [1]. The term lupus coined initially in the Middle Ages to portray erosive skin lesions reminiscent of a 'wolf's bite'. Viennese physicians Ferdinand von Hebra initiate the butterfly allegory to depict the malar rash (found in the skin) as well used the term 'lupus Erythematosus' and published the primary illustrations in his Atlas of Skin Diseases in 1856. Lupus was first predicted as a systemic disease with instinctive manifestations by Moriz Kaposi in 1837–1902. The complete form was further recognized by Osler in Baltimore and Jadassohn in Vienna. Other vital milestones include the portrayal of the false positive test for syphilis in SLE by Reinhart and Hauck from Germany in 1909; the description of the endocarditis lesions in SLE by Libman and Sacks in New York in 1923.

A. National statistics

The prevalence of lupus in India is approximately 5,481,981. Most of the Indians are not aware of this disease. Hospitals are lacking due to less optimal management facilities. Diagnosis is not done in earlier stage due to lack of awareness about this disease. 50% of the patients consult at least three physicians and specialist for a minimum of 4 years before the diagnosis of lupus. After knowing the severity and seriousness of the disease, patients are approaching the city hospitals. These data are mentioned in centres for Disease Control and Prevention (CDC).

B. International statistics

The maximum rate of commonness has been reported in Spain, Italy, Martinique, Afro-Caribbean and the United Kingdom. Although the occurrence of SLE is huge in black persons in the United Kingdom, the disease is hardly reported in African blacks, signifying an ecological trigger, as well as a genetic basis, for disease in the UK population [2]. The annual incidence of SLE averages 5 cases per 100,000 populations. The Centres for Disease Control and Prevention (CDC) estimates a range between 1.8 and 7.6 per 100,000 persons per year in the continental United States. The Lupus Foundation of American assessments states the occurrence to be up to 1.5 million cases, which likely reflects the annexation of minor forms of this disease. The occurrence of SLE varies by ethnicity and ethnicity where higher rates of possibilities reported in blacks and Hispanics. The occurrence of SLE in black women is nearly 4 times higher than that in white women. SLE is also more frequent in Asian women than in white women.

2. Applying data mining to predict lupus

Usually the hospitals will be maintaining Databases which have query facility. The normal querying tools yield the solution for queries such as longer stay due to surgeries, health status of a patient, next appointment date, and so on but data mining can answer more convoluted and valuable queries like the type of disease, comparison with the previous related records and learning through the existing records is possible with data mining. Additionally data mining in health care is also used for efficient data management, effective data processing and so on. Clinical

determinations are mostly made based on doctors' suspicion and occurrence rather than on the knowledge rich data concealed in the database. Thus leads to outcast increased medical cost, biases and errors which affects patients' health more and affects the quality of service provided to patients and affects the fame of the hospital.

In health care, data mining is becoming increasingly popular, as also essential. Data mining applications can greatly benefit doctors, medical practitioners, nurses, organizations and those who are involved in the health care industry [3]. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision-making. This article explores data mining applications for the early prediction of SLE disease. In particular, it discusses an enhanced algorithm and its applications for the early prediction of the autoimmune disease. Finally, the research highlights the causes of lupus and the method to predict the disease in advance and to extend the life span of the patients. By processing the data with K-means [4, 5] and J48 algorithm [6, 7] has castoff to predict lupus in advance.

3. Origin of the research problem

The human population is increasing day by day as also diseases are spreading fast and new diseases are occurring all over the world. No permanent cure has been found for some diseases. One of the diseases which have no permanent cure is Systemic Lupus Erythematosus also called as Lupus or SLE. Lupus affects mostly Americans, Africans, Asians, Hispanics, and Caucasians, now it's commonly occurring in India. The origin of SLE from India is reported on 1995 and it extends to 1366 patients with lupus till 1998. The latest details about the affected persons are still under research as the people are not aware of the disease [8, 1]. The survey says that 29% of women could not define lupus and 31% of the populations are not aware that child and women are at risk and there is no awareness about the disease.

A. Problem specification

Since Lupus is an autoimmune chronic disease which pretends to be like many other diseases, it is challenging to diagnose in the initial stage. Lupus cannot be cured and there is no specific medicine, treatment or surgery to cure the disease. The only solution is to extend the life span of the patients if they are diagnosed in the early stage. The center of disease control states that high causes of lupus in India are due to less optimal facilities and lack of awareness among the public. The cause of lupus is unknown and leads to the late prediction of the disease.

B. Research objective

The main objectives of the research are -

- To study and analyze the lupus patients real-time data set
- To design and propose a new HKMJDT algorithm to predict the disease in early stage.
- To analyze the various algorithms to spectacle the best algorithm which will be more effective for the prediction
- To extend the life span of the Lupus patient.

4. Literature review

Sayad et.al., [9] designed a decision support system to predict heart disease using ID3 algorithm and multilayer perceptron with back propagation as training algorithm. This 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 effectively. Discovering the hidden patterns and relationship between them is often unexploited. The data was collected from Cleveland University and the results were tabulated. A sample decision tree, system architecture and analysis result of three algorithm is summarized.

The clinical profile of Systemic Lupus Erythematosus (SLE) patients at a tertiary care in western India has been published by the Saigal et.al., American College of Rheumatology (ACR) suggested 11 criteria to diagnose lupus disease [10]. ACR criteria are considered to be the important analysis criteria. Authors evaluated 60 lupus patients data set over a period of one year. Arthritis was found as common manifestation among SLE patients. Other symptoms are cutaneous, cardiac, renal, neuropsychiatric, gastro intentional etc. Lupus is multi-system disorder and common among females. The analysis report of 60 patients were tabulated. The appropriate management of lupus is critically depending on the proper assessment of disease activity, quality of life and organ damage [11].

Assessment of lupus is not based on single test and it involves accurate physical and laboratory diagnosis, recording of accurate morbidity, monitoring of disease activity and integration of these with the patient's own perceptions of health status and quality of life. The etiological factor of lupus is multi-factorial; the disease is characterized by the production of auto-antibodies which leads to immune complex deposition, inflammation and eventually permanent organ damage. Various disease index like Systemic Lupus Erythematosus Disease Activity Index (SLEDAI), British Isles Lupus Assessment Group (BILAG), European Consensus Lupus Activity Measurements (ECLAM) and Systemic Lupus Activity Measure (SLAM) are also tabulated with the various measures and scores to assess and diagnose lupus efficiently. Lupus is characterized by periods of relative quiescence and periods of exacerbations which involve any organ or system in various combinations of body was analysed [12].

The SLE patients are affected by permanent organ system damage. The damage will progress over time and will be severe in African blacks than in American whites. The prognosis of lupus has improved over last 4 decades.

5. Data analysis

Data obtained from the lab has been preprocessed using Pentaho Data Integration tool [13]. The disease will pretend to be like many other diseases thus the patients will have many symptoms and the symptoms are related to many other diseases. The preprocessed data doesn't have any labeled field and the proper prediction cannot be obtained by analyzing the unlabeled data. The various symptoms which are common to other diseases are shown in the TABLE I. Hence direct prediction of disease with classification algorithm is impossible. The data need to be clustered based on the distance. The cluster should be formed first and based on the cluster, the data need to be classified using classification algorithm. Combining K-means clustering algorithm is used by Bouhmala et.al.,[14] in the research. Table II tabulates the list of attributes considered for prediction.

Table 1: Lupus Shows the common symptoms which can also be wrongly predicted as some other disease. This table shows the reason for clustering the data before predicting the disease directly.

Symptoms	Diseases
Extreme fatigue (tired-ness)	Lupus, Cancer, AIDS, Dengue, Diabetes, TB
Painful or swollen joints	Lupus, AIDS, Dengue
Headaches	Lupus, Cancer, AIDS, Dengue
Fever	Lupus, Dengue, Cancer, TB
Anemia (low RBC, WBC)	Lupus, Diabetes, TB

Table 2. Shows the description of the dataset set used for this research. The data set have been obtained from the lab. The raw data set has many fields like patient father name, patient spouse name etc. The data has been pre-processed and the essential attributes are tabulated

Attribute No	Attribute Name	Attribute Description	Values
1.	PatientID	Patient I is to unique	P001
2.	SampleType	Sample type from the patient for lab	Serum/plasma/urine
3.	Volume	Quantity of sample type	15 ml
4.	Age	Patient's age	1-75
5.	Gender	Patient's gender	0:Male 1: Female
6.	Ethnicity	Patient's Ethnicity	0:Indian, 1:Caucasian 2: American, 3: Hispanic, 4:African
7.	Diagnosis	Patient's affected body parts	0: Seizue, 1: Psychosis 2: Brain Syndrome, 3: visual disturbance 4: Carnial nerve disorder 5: Lupus headache 6: Cerbovascular accidents 7: Vasculities, 8: Arthritis 9: Myositis, 10:urinary cast 11: hematuria, 12: Protenuria 13:pyonuria, 14:New rash 15: Alopecia, 16: mucosal ulcer 17: Pleurisy, 18: Pericarditis 19: Low complement 20: Increased DNA binding 21: Fever, 22: Thrombocytopenia 23: Leucopenia
8.	SLEDAI Score	Score based on diagnosis	1-21
9.	AgeAtDiagnosis	Patient diagnosed as he/she has lupus	1-75
10.	Disease Activity	Disease stage	0: SLEDAI Score 0 - No Activity, 1:SLEDAI Score 1-5- Mild Activity, 2: SLEDAI Score 6-10 – Moderate, 3: SLEDAI Score 11-19 – High, 4: SLEDAI Score 20 – very high
11.	Most Recent Lab result	Lab result of patient	0: ANA, 1:Anti dsDNA 2: sm, 3: CRP 4: Urine test, 5: PT, 6: PTT
12.	Profession	Patient's nature of job	0:Factory work- exposure of silica 1:Night shift (Non exposure of Vitamin D), 2: Pesticides 3: Pollutants, 4: Smoke area
13.	Smoking Habit	Patient's habit of cigarette	0:Yes, 1: No 2: Occasional

6. Statistical analysis

The statistics measure has been calculated before clustering and after clustering with k-means algorithm to show the reason of implementing k-means clustering in this prediction process. The data has been evaluated with the classification algorithms J48, Hoeffding tree, Naïve-Bayes, random table and decision tree algorithms before and after clustering with K-means algorithm to identify the best combination of the algorithms. The enhanced algorithm is designed based on the best statistical results. TABLE II tabulates the list of attributes considered for prediction. TABLE III shows the confusion matrix. Confusion matrix is a special table layout that lets visualization of the performance and effectiveness of an algorithm, typically for a supervised learning (classifier). The row represents the instances in a predicted class and column represents the instances in actual class.

Table 3: Confusion Matrix

	No of records = 112	**True outcome: (Patients have lupus disease)**	
		P (Patients have Lupus disease)	**N (Patients do not have lupus disease)**
Actual Class	**P (Patients have Lupus)**	**TP** (Patients Correctly predicted as Lupus)	**FP** (Patients who have Lupus wrongly predicted as they are normal)
	N (Patients do not have lupus)	**FN** (Patients who are normal but predicted with lupus)	**TN** (Patients who are normal, predicted to be normal)

6.1. Accuracy

The accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's true value (i.e.) the patients who are approximately predicted to have lupus. Accuracy is the proximity of measurement results to the true value; precision, the repeatability, or reproducibility of the measurement. Fig. 1 shows the accuracy comparison chart.

$$AC = \frac{\sum TPTN}{\sum TPTNFPFN} \quad (1)$$

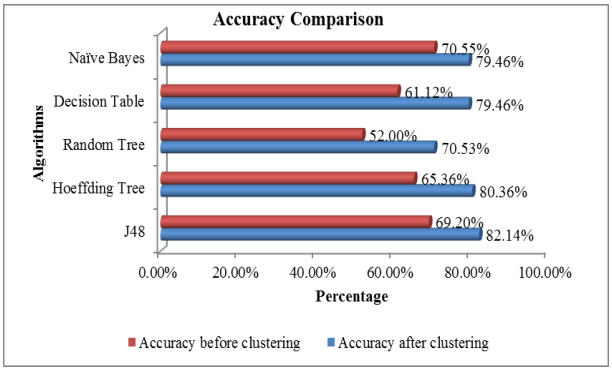

Fig. 1: The accuracy comparison of algorithm is depicted. Accuracy before clustering the data set is low compared with the accuracy of clustered data. The clustering is done with K-means algorithm. When k-means is combined with various classification algorithm, it shows better results with J48 algorithm

6.2. Specificity

Specificity relates to the test's ability to correctly detect patients without a condition. Specificity of a test is the proportion of healthy patients known not to have the disease, who will test negative for it. Fig 2 shows the specificity comparison chart.

$$SP = \frac{TN}{TN+FP} \quad (2)$$

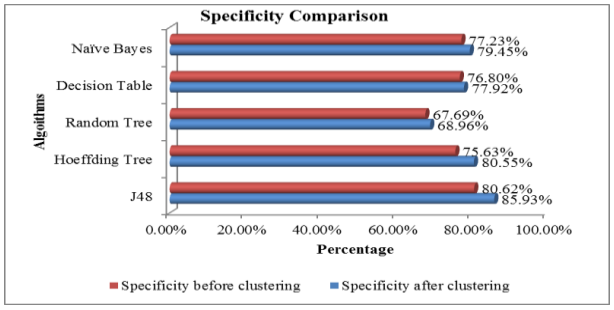

Fig. 2: The Specificity comparison is shown. The diagram shows the specificity of various algorithms before and after clustering the dataset. K-means clustering has been used to cluster the dataset. J48 with K-means shows the best specificity result compared with other classification algorithms.

6.3. Sensitivity

Sensitivity refers to the test's ability to correctly detect patients who do have the condition. A negative result in a test with high sensitivity is useful for ruling out disease. Fig 3 shows the sensitivity comparison chart.

$$SN = \frac{TP}{TP+FN} \quad (3)$$

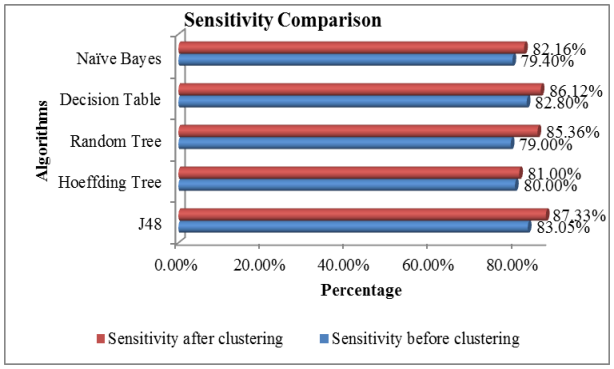

Fig. 3: The Sensitivity of the dataset when applied after clustering with K-means algorithm shows the best result when compared with other classification algorithm. The data set before clustering shows less percentage when compared with the clustered data set.

6.4. Precision

Precision is the randomly selected or retrieved data is relevant. Precision is also called as Positive Predictive Value (PPV). The random selection should be such that all data in the data base are equally likely to be selected. Fig 4 shows the precision comparison chart.

$$PREC = \frac{TP}{TP+FP} \quad (4)$$

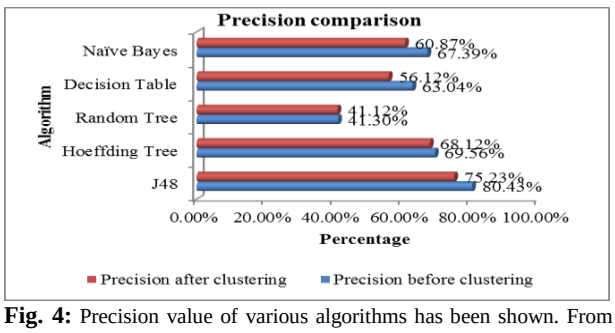

Fig. 4: Precision value of various algorithms has been shown. From the picture it is clear that the precision value is comparably high when the data set is clustered and classified with k-means and J48 algorithm.

6.5. False Omission Rate

False omission rate (FOR) is a statistical method used in multiple hypothesis testing to correct for multiple comparisons and it is the complement of the negative predictive value. It measures the proportion of false negatives which are incorrectly rejected. Fig 5 shows the FOR comparison chart.

$$FOR = \frac{FN}{FN+TN} = 1 - NPV \quad (5)$$

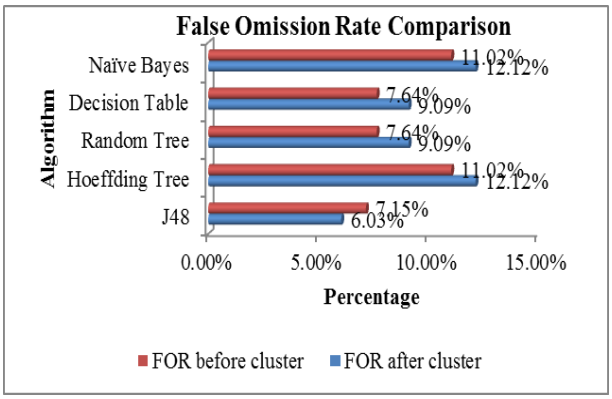

Fig. 5: The False Omission Rate (FOR) of various algorithm have been shown in the above picture. The clustered data set with J48 algorithm shows the best result. K-means with J48 shows the less False Omission rate which can be calculated with the confusion matrix Hybrid k-means J48.

7. Decision Tree Algorithm (HKMJDT)

Algorithm: HKMJDT algorithm

```
1:    procedure HKMJDT ( )
// R={r1,r2,….,rn} (set of records to be clustered);
// K (number of cluster); MaxLoop (Limit of loop Iterations);
// N (Node);
// T (Training Data);
// AA (Attributes_Available);
// BA (Best_Attribute);
// C={c1,c2,….ck}  (set of cluster Centroids);
// L={l(r) | r = 1,2,…,n}  (set of cluster labels of R);
// LN (LeafNode)
2:    begin
3:    foreach ci ∈ Cdo
4:    ci  = rj ∈ R      // Random selection
5:    end
6:    foreach ri ∈ R do
7:    l(ri) = EucledianDistance √((ri − c1)² + (rj − c2)²) ∈ (1,2, … , k}
8:    end
9:    flag = false; loop = 0;
10:   repeat
11:   foreach ci  ∈  C do
12:                 updateCluster(ci)
13:    end
14:    foreach ri ∈  R do
15:                 minDist          EucledianDistance √((ri − c1)² + (rj − c2)² ∈ (1,2, … , k}
16:                 if  (minDist f= l(ri))
17:                       l(ri) = minDist;
18:                       flag = true;
19:                 end;
20:                 end;
21:                 loop++;
22:       Until flag = true and loop
23:                       MaxLoop
24:     Using L, Create N
25:     if  all records in T  is same as L
26:                 return n as LN ;
27:     end
28:     If  AA is empty
29:                  return n as LN          // with maximum
```

```
30:     end
31:     if  BA(T, AA)
32:                 AA = AA - BA
33:                 SplitRecords BA(BA,T)
34:                 foreach split build
35:                     OCBC ( )
36:                     foreach split Ti of T on BA
37:                             Add new node to OCBC
38:                     end;
39:                 end;
40:   end;
```

8. Workflow of the Proposed Work

Fig 6. depicts the workflow of the proposed algorithm. The raw dataset obtained from the lab has been analysed and the selected data has been considered for further process. The selected data set has been pre-processed and used for the prediction. The data is clustered with k-means clustering initially and the final prediction is made by the J48 algorithm. Two clusters will be formed one for patients with lupus and the other with normal conditions. In clustering process, it randomly chooses K objects and makes them the K cluster centroids. The distance between each cluster centroid and the record has been calculated. The records that have minimum distance will be assigned to the nearest clusters, recalculate the cluster means and check till the last record in the database. Once the cluster has been formed, label the clusters.

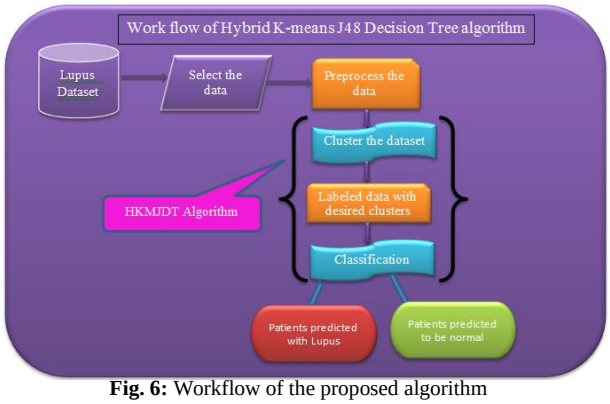

Fig. 6: Workflow of the proposed algorithm

With that label, the decision tree will be formed and the patients who are predicted with lupus and the patients who don't have lupus will be predicted. The J48 algorithm is an open source Java implementation of C4.5. The J48 algorithm uses both C4.5 confidence based post-pruning and sub tree-raising. Derivation of rules, decision tree pruning, and continuous attribute ranges and accounting for missing values are the major features of J48 algorithm.

9. Conclusion and future work

Naïve Bayes, J48, Hoeffding tree, random tree and decision table are implemented on the input data before and after clustering with K-means algorithm to assess the best performing algorithm. The present work uses Hybrid K-means J48 Decision Tree algorithm to obtain the best accuracy for the early prediction of SLE. The reason for choosing two algorithms is described with necessary charts. The statistical measures like accuracy, specificity, sensitivity, precision and false omission rate to identify the suitable algorithm is also been discussed. The future work is to develop a "Lupus Prediction System" with the proposed algorithm to predict the disease. Before implementation, suggestions will be obtained from various doctors and medical practitioners. The public, who needs to know about the Lupus and its severity, can just enter the symptoms and know the current stage of the disease.

References

[1] Gill JM, Quisel AM, Rocca PV, Walters DT. Diagnosis of systemic lupus erythematosus. J American family physician 2003; 12: 2179-2186.

[2] Ben-Menachem E. Systemic lupus erythematosus: A review for anesthesiologists. J Anesthesia & Analgesia 2010; 9: 111(3):665-76.

[3] [3] Wakoli, Leonard Wafula, Abkul Orto, and Stephen Mageto. Application of The K-Means Clustering Algorithm In Medical Claims Fraud/Abuse Detection. International Journal of Application or Innovation in Engineering & Management 2014; 7: 142-151.

[4] [4] Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence 2002; 7: 881-892.

[5] Li T, Bai S, Ning J. K-means, an applicable and efficient clustering algorithm. Energy Procedia 2011; 11: 3189-3196.

[6] Kaur G, Chhabra A. Improved J48 classification algorithm for the prediction of diabetes. International J of Computer Applications. 2014; 7: 13-17.

[7] Wagh S, Khati A, Irani A, Inamdar N, Soni R. Effective Framework of J48 Algorithm using Semi-Supervised Approach for Intrusion Detection. International Journal of Computer Applications 2014; 5: 23-27.

[8] Carreno LJ, Pacheco R, Gutierrez MA, Jacobelli S, Kalergis AM. Disease activity in systemic lupus erythematosus is associated with an altered expression of low-affinity Fcγ receptors and costimulatory molecules on dendritic cells. J Immunology 2009; 11: 334-41.

[9] Sayad, A. and Halkarnikar, P. Diagnosis of heart disease using neural network approach. In Proceedings of IRF International Conference, 13 April 2014; Pune, India: pp. 978–993.

[10] Saigal, R., Kansal, A., Mittal, M., Singh, Y., Maharia, H. R. and Juneja, M. Clinical profile of systemic lupus erythematosus patients at a tertiary care centre in western india. J Indian Acad Clin Med 2011; 1: 27–32.

[11] Lam GK, Petri M. Assessment of systemic lupus erythematosus. J Clinical and experimental rheumatology 2005; 9: 20-132.

[12] Gladman DD, Urowitz MB, Esdaile JM, Hahn BH, Klippel J, Lahita R, Liang MH, Schur P, Petri M, Wallace D. Guidelines for referral and management of systemic lupus erythematosus in adults. J Arthritis and Rheumatism 1999; 9:1785-1796.

[13] Gomathi, S. and Narayani, V. Preprocessing systemic lupus erythematosus (sle) data set with pentaho data integration (pdi). International J of Recent Innovation in Engineering and Research 2017; 3: 75-79.

[14] Bouhmala N, Viken A, Lonnum JB. Enhanced Genetic Algorithm with K-Means for the Clustering Problem. Int J of Modelling and Optimization 2015; 3: 150-154.