Deep Insight to Implement Big data analytics to predict Systemic Lupus Erythematosus (SLE)

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

Assistant Professor, IT, SKASC, Research Scholar, Bharathiar University, Coimbatore, India [1]
Director I/C, Department of MCA, Karpagam College of Engineering, Coimbatore, India [2]

***Abstract*—Medical data comprise of massive formats of information such as images, audio, video, reports etc. Handling all these data need advanced technique and technology which must have large storage, centralized distribution and patients record management which is highly possible with big data. Big data is majorly used in clinical operations, research and development and public health care informatics. The prediction of autoimmune disease like Lupus will be made easier with big data since it needs many lab test reports and imaging reports. The main objective of the paper is to show the methodology to implement big data to predict lupus.**

***Index Terms*—Data mining, big data, lupus, autoimmune disease, hadoop, classification, SLEDAI, ACR, classification.**

I. Introduction

A. *Systemic Lupus Erythematosus*

Systemic lupus Erythematous is a multisystem autoimmune disorder which is complex to diagnose at the early stage. The disease spoils more than two organs thus difficult to diagnose with single test and which can be analyzed through the combination of clinical and laboratory tests. American college of Rheumatology suggested eleven criteria to diagnose lupus. Mostly SLE affects more female than male in the ratio of 4:1 [13] and most commonly affect blacks, Hispanic and Asian races than whites. There are no common and possible symptoms among the lupus patients. Causes and tests may vary based on the problems. The variability of individual causes among patients leads to complications which lead to lack of diagnosing the disease earlier. There is no medicine or vaccine to cure the disease but the life time of the patients can be extended if the disease is predicted in initial stages [14].

The analysis of the disease require EMRs, financial & operational data, clinical data, personal medical records, clinical trials, radiology images, 3D imaging, biometric sensor etc. These set of data need large storage and processing which can be effectively established using big data. The various sources to diagnose lupus disease include prescription, physician written notes, medical images; pharmacy notes etc. These data are necessary to discover associations and to understand patterns & trends within the data. Big data analytics is a powerful way to diagnose the disease, care the life of patients and extend their life time. Big data deals with the huge variety of data with large volume and velocity.

B. *Big Data*

Outmoded medical practice is moving from relatively ad-hoc and idiosyncratic decision making to modern evidence-based healthcare. Evidence is based on data collected from electronic genomic information; health record (EHR) systems; sensors, capturing devices, and mobiles; information transferred verbally by the patient; and new medical awareness. These resources yield a throng of hefty and intricate datasets, which are challenging to process using common database management tools or traditional data processing applications.

Currently, the data spawned in the method of medical care in intensive care units (ICUs) are rarely handled in real-time, nor are they collected and used for data analysis. The actuality of user-friendly platforms for retrieving and exploring such substantial volumes of data could influence an eon of medical knowledge discovery and medical care quality enhancement. The current absence of such platforms is due partly to the difficulty of accessing, unifying, and corroborating baggy health care data produced with high time-varying arrival rates by various types of devices, and especially to the jaggedness of proprietary software solutions in use. The need for fast computationally-intensive systematic exploration of the data engendered in ICUs obstructs the use of traditional databases and data processing applications and demands the development of cloud based solutions.

Data undergoes three stages (data capture, provisioning and data analysis) before it can be used for sustainable, meaningful analytics which is shown in Fig 1. Data capture shows how does data get into your system? The solution for the query is Data Provisioning is addressed based on the cost, quality, EMR data for lab outcomes, clinical observations, ADT information. Billing those data is to help ascertain cohorts using diagnosis and technique codes and charges and revenue, Cost data (if available) to view the improvement's impact on margins, Patient Satisfaction data and patient satisfaction [1]. Data analysis deals with four tasks which includes Data quality evaluation, Data Discovery, Interpretation and Presentation.

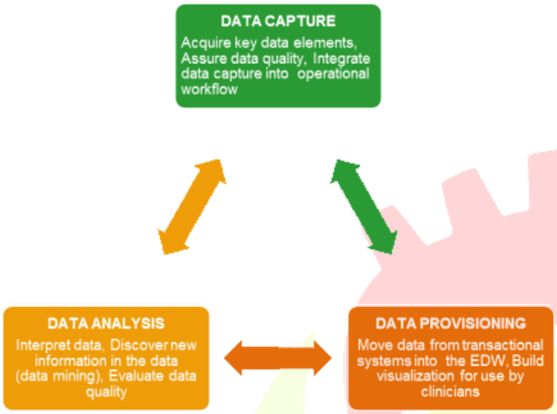
DATA CAPTURE
Acquire key data elements, Assure data quality, Integrate data capture into operational workflow
DATA ANALYSIS
Interpret data, Discover new information in the data (data mining), Evaluate data quality
DATA PROVISIONING
Move data from transactional systems into the EDW, Build visualization for use by clinicians

Fig. 1. Analyzing Healthcare data

II. Problem Specification

1) Traditional technique cannot predict the disease in initial stage, thus big data is implemented to detect the disease earlier and to provide efficient treatment.

2) Length of Stay (LOS) can be known in advance.

3) Cost effective way to diagnose and treat patients.

III. Literature Review

Garlasu, D et.al., penned about Grid Computing shows the improvement about the stowage proficiencies and the handling power and the Hadoop technology is used for the enactment tenacity.

The advantage of Grid computing center is the high stowing capability and the high processing power. Grid Computing makes the immense contributions among the methodical exploration, help the scientists to examine and hoard the large and intricate data [5].

Mukherjee. A et.al., describes big data analytics define the exploration of large amount of data to acquire the expedient information and reveal the veiled patterns. Big data analytics refers to the Mapreduce Structure which is developed by the Google. Apache Hadoop is the open source platform which is used for the purpose of enactment of Google's Mapreduce Model [2].

Ramsey-Goldman and Manzi [1] have recently shown a connotation between lessened bone mineral density (BMD) and both an augmented carotid plaque index and the existence of coronary artery calcification of pilot study of 65 women who affected with lupus. This supports the notion that inflammatory and immune-mediated mechanisms involved in lupus may also subsidize to the development of atheroma and osteoporosis.

Kipen *et al*. [3] studied 97 female lupus patients with a mean age of 44.2□yrs and found that there was low bone mass (>1□S.D. below young adult mean) in the spine and femoral neck in over 40% of the patients. There was osteoporotic level BMD (>2.5□S.D. below the young adult mean) in the spine of 13% of the patients and in the femoral neck of 6% of the patients [3]. There was a much clearer inverse relation between steroid use ever and the spine BMD result than the femoral neck BMD.

In Birmingham we have studied 242 patients, median age 39.9□yr (range 18–80□yr) [4]. We found that 10% of our patients were osteoporotic and 41% were osteopenic by BMD scanning. Fractures had occurred in 9% of patients since the onset of lupus in the absence of significant trauma; one in five of those who were osteoporotic, one in seven of those who were osteopenic and one in 22 of those with normal BMD at spine and femoral neck.

IV. Proposed Work

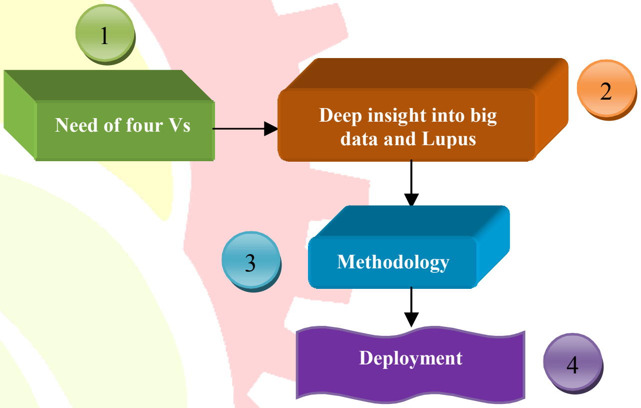

Fig. 2. Framework

The proposed work is divided into four steps. The very first Step is the nasic one which involves in identifying the needs of four Vs. The disease is identified with the various parameters, that is tabulated in Table 1. Big data deals with the three types of data namely structured, unstructured and semi structured data.

TABLE I: Big data to predict Lupus

Type of Data	Meaning	Lupus data
Structured	Normal Human readable form of data. Eg. Text, Doctors hand written report	Electronic Health Records (EHR), Patients detail, bills
Semi structured	Irregular structure which is not known in advance	XML data while transferring the patients detail through internet.
unstructured	Does not follow any format or rules and unpredictable.	CBC result, charts, Urinalysis, ANA test report, APL report, CRP report

Table II shows the detailed description of four stages.

TABLE II: Framework explanation

Four Vs.	**Volume, Velocity, Veracity, Variety.**
Deep Insight into Big data &Lupus	**a) Problem being addressed** **b)importance and interesting facts** **c)Need of big data** **d)Data sets**
Methodology	**a)Sub problems** **b)Variable selection from data set** **c)Data collection** **d)ETL & data transformation** **e)Tool selection** **f) Conceptual model** **g)Analytic technique** **h)Mining technique** **i) Results**
Deployment	**a) Evaluation and Validation** **b)Testing**

The various signs and symptoms of SLE involves fatigue, low grade fever, loss of appetite, weight loss, arthritis, Myalgia, Malar rash, photosensitivity, pleuritis, pericarditis, Raynaud's phenomenon, alopecia, jaundice, lymphedema. The identification and prediction is more intricate since it imagines to be like many other disease. American College of Rheumatology (ACR) suggested 11 criteria to diagnose lupus. If 4 criteria are present in the patient then he/she is assumed to be affected with Lupus.

TABLE II: 11 ACR CRITERIA

S. No	Criteria	Explanation
1	Malar Rash	Fixed erythema, flat or raised, over the malar eminences
2	Discoid Rash	Erythematous raised patches
3	Photosensitivity	Exposure to UV light causes rash
4	Oral Ulcers	Includes oral and nasopharyngeal, observed by physician
5	Arthritis	Non erosive arthritis involving two or more peripheral joints
6	Serositis	Pleuritis or pericarditis documented by ECG or rub or evidence of pericardial effusion.
7	Renal disorder	Proteinuria > 0.5 g/d or > 3+, or cellular casts
8	Neurologic disorder	Seizures without other cause or psychosis without other cause
9	Hematologic disorder	Hemolytic anemia
10	Immunologic disorder	Anti-dsDNA, anti-Sm, and/or anti-phospholipid
11	Antinuclear antibodies	An abnormal titer of ANAs by immunofluorescence

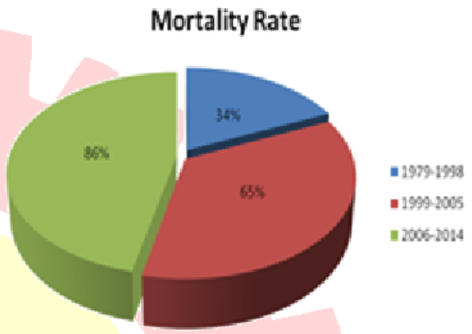

Fig. 3. Mortality rate of SLE

A. *Evaluation Criteria of Lupus (SLEDAI)*

There are two major scoring systems to evaluate the activity of lupus in clinical studies. These systems are not used in routine medical practice, but for quantification of lupus disease activity primarily for the purpose of determining whether a new drug evaluated for the disease is effective[20].

The most commonly used study of lupus activity is called the SLE Disease Activity Index, and the acronym for it is SLEDAI. It is a list of 24 items, 16 of which are clinical items such as seizure, psychosis, organic brain syndrome, visual disturbance, other neurological problems, hair loss, new rash, muscle weakness, arthritis, blood vessel inflammation, mouth sores, chest pain worse with deep breathing and manifestations of pleurisy and/or pericarditis and fever.

Eight of the 24 items are laboratory results such as urinalysis testing, blood complement levels, increased, low platelets, anti-DNA antibody levels and low white blood cell count. These items are scored based on whether these appearances are extant or vague in the previous 10 days [17].

Organ involvement is weighted; for example, kidney disease, joint pain and are each burgeoned by four, but central nervous system neurological involvement is multiplied by eight. The slanted organ manifestations are then summed into a final score, which can range from zero to 105. Scores greater than 20 are rare [17, 18]. A SLEDAI of 6 or more has been shown to be consistent with active disease requiring therapy. A clinically evocative modification has been conveyed to be an enhancement of 6 points or worsening of 8 points.

The SLEDAI was reformed in the Safety of Estrogens in Lupus Erythematosus National Assessment (SELENA) trial; this modification is known as the SELENA-SLEDAI system. The SELENA-SLEDAI adds some clarity to some of the definitions of activity in the discrete items, but does not change the rudimentary notching system [18].

The further foremost study mechanism is called the BILAG, which stands for the British Isles Lupus Activity Group. The BILAG is an organ-specific 86- query assessment established on the standard of the doctor's committed to treat, which entails an assessment of improved the similar, worse, or new over the last month. .

The ensuing scores for every organ can be A through E, where (1) A is very active disease, (2) B is moderate activity, (3) C is mild stable disease, (4) D is resolved activity, and (5) E indicates the organ was never involved [16]. There are 8 broad captions; Mouth and Skin, General, Joints and Muscles, Neurological, Pulmonary and Cardiovascular, Blood Vessel Inflammation (Vasculitis), Blood and Kidney.

B. *Big data to predict Lupus*

A report delivered to the US congress in August 2012 describes big data as "High velocity and large volumes of complex data with variety of data which require advanced techniques and technology to make efficient storage, wide distribution and effective management and analysis of information". Lupus patients data set deals with variety and high velocity of data thus can be effectively dealt with big data for better predictions. Fig. 3 shows the mortality rate of lupus which is high during 2006-2014.

C. *Classification algorithm to predict lupus*

Methods: Attribute Selection method, splitting_attribute
Algorithm: Generate_decision_tree
Input:
Step 1: Data partition, D, which is a set of training tuples and their associated class labels.
Step 2: Attribute_list, the set of candidate attributes.
Step 3: Attribute selection method: a procedure to determine the splitting criterion that best partitions the data tuples into individual classes. This criterion includes a splitting_attribute and either a splitting point or splitting subset.
Output:
A Decision Tree

```
Process
create a node N;

if tuples in D are all of the same class, C then
  return N as leaf node labeled with class C;

if attribute_list is empty then
  return N as leaf node with labeled
  with majority class in D;|| majority voting

apply attribute_selection_method(D, attribute_list)
to find the best splitting_criterion;
label node N with splitting_criterion;

if splitting_attribute is discrete-valued and
  multiway splits allowed then
 // no restricted to binary trees

attribute_list = splitting attribute;
 // remove splitting attribute
for each outcome j of splitting criterion

  // partition the tuples and grow sub trees for each
partition
  let Dj be the set of data tuples in D satisfying
outcome j; // a partition

  if Dj is empty then
    attach a leaf labeled with the majority
    class in D to node N;
  else
    attach the node returned by Generate
    decision tree(Dj, attribute list) to node N;
  end for
return N;
```

The procedure to classify the data set is shown and the detail technique is also discussed in this section. This procedure will yield a decision tree. This algorithm is implemented in Hadoop for fast and efficient output.

D. *Hadoop*

Hadoop is built with Apache unspoilt source framework coded in java that consents scattered handling of hefty datasets athwart clusters of computers using simple coding prototypes. A Hadoop frame-worked application works in a milieu that provides disseminated stowage and computation across clusters of PCs [20]. Hadoop is intended to scale up from solitary server to thousands of technologies, each offering local reckoning and storage.

V. Conclusion

Big data analytics is an efficient technology which is widely used in many sectors. Medical sector is one of the highlighted sectors where the data are incredible. Thus applying big data in health care will yield an efficient outcome and timely decision making. This paper highlights some major importance of using big data to predict Systemic Lupus Erythematosus. The future work will be to apply the big data with the decision tree to predict the lupus disease.

References

(Periodical style)

[1] Kipen Y, Buchbinder R, Forbes A, Strauss B, Littlejohn G, Morand E. Prevalence of reduced bone mineral density in systemic lupus erythematosus and the role of steroids. J Rheumatol 1997;24:1922–9

[2] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , “Shared disk big data analytics with Apache Hadoop”.

[3] Sinigaglia L, Varenna M, Binelli L *et al*. Determinants of bone mass in systemic lupus erythematosus: a cross sectional study on premenopausal women. J Rheumatol 1999;26:1280–4.

[4] Gordon C, Crabtree N, Skan J, Bowman S, Situnayake D. Prevalence and predictors of osteoporotic fractures in patients with systemic lupus erythematosus. Arthritis Rheum 2001;44(Suppl):S334..

[5] Garlasu, D.; Sandulescu, V. ; Halcu, I. ; Neculoiu, G. ;,(17-19 Jan. 2013),”A Big Data implementation based on Grid Computing”, Grid Computing.

[6] Bloch DA, Olshen RA, Walker MG (2002) Risk estimation for classification trees. J Comput Graph Stat 11:263–288

[7] Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inform Theory 13(1):21–27

[8] Domingos P (1999) MetaCost: A general method for making classifiers cost-sensitive. In: Proceedings of the fifth international conference on knowledge discovery and data mining, pp 155–164.

[9] Kuramochi M, Karypis G (2005) Gene Classification using Expression Profiles: A Feasibility Study. Int J Artif Intell Tools 14(4):641–660.

[10] Yin, X. & Han, J. CPAR: Classification based on predictive association rule. In Proceedings of the SIAM International Conference on Data Mining. San Francisco, CA: SIAM Press, 2003,pp. 369-376.

[11] Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan et al. "Top 10 algorithms in data mining." Knowledge and Information Systems 14, no. 1 (2008): 1-37.

[12] Han Jing-ti and Gu Yu-jia (2009) Study on Handling Range Inputs Methods On C4.5 algorithm. IEEE International Forum on Computer Science – Technology and Application.

[13] Edworthy SM. Clinical manifestations of systemic lupus erythematosus. In: Ruddy S, Harris ED, Sledge CB, Kelley WN, eds. Kelley’s Textbook of rheumatology. 6th ed. Philadelphia: Saunders, 2001:1105-19.

[14] Bellomio V, Spindler A, Lucero E, Berman A, Santana M, Moreno C, et al. Systemic lupus erythematosus: mortality and survival in Argentina. A multicenter study. Lupus 2000;9:377-81.

[15] Schur PH. General symptomatology and diagnosis of systemic lupus erythematosus in adults. Retrieved March 20, 2003, from http://www.uptodate.com/physicians/rheumatology_toclist.asp.

[16] http://www.lupusil.org/what-are-the-sledai-and-bilag-evaluations.htm l

[17] http://www.ncbi.nlm.nih.gov/pubmed/1599520

[18] http://www.ncbi.nlm.nih.gov/pubmed/11838846

[19] http://rheumatology.oxfordjournals.org/content/early/2011/01/17/rheumatology.keq376.full

[20] http://www.tutorialspoint.com/hadoop/hadoop_introduction.htm