



## Application Of Data Mining In Detecting Pattern Of Disease Spread In Various States Of India

**Hemlata Sharma<sup>1</sup>**Senior Assistant Professor-IT,  
LBSIM, New Delhi,  
India**Pallavi Sharma<sup>2</sup>**Senior Assistant Professor-IT,  
LBSIM, New Delhi,  
India

---

**Abstract**— *Data mining has been used intensively and extensively by many organizations. In healthcare, data mining is becoming increasingly popular. These organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services. By using Data Mining Techniques like classification and clustering author finds hidden patterns which give meaningful decision making to epidemic diseases and the impact of diseases in the various states of India.*

**Keywords :** Hepatitis, Dengue, Acute respiratory, Kala-azar, Encephalitis ,Clustering and Classification .

---

### I. INTRODUCTION

The new upcoming area in IT industry is health care . Due to huge amount of data available in health care, it has become more difficult to identify the usable and important data for the prediction, prevention and control of diseases. By using Data mining we can resolve these issues, but for this it requires a strong collaboration of health care and IT experts to work on same platform.

In our study we have analyzed health data of various states of India to identify the impact of mortality rate due to major diseases in various zones and to find out the similarity of patterns in between these diseases which will benefit the society in many ways like by using it Physicians may identify effective treatments and best practices, and Patients receive better and more affordable healthcare services.

This paper is broadly categorized into the following sections: I Introduction : under this section we showcased the objectives and the whole structure of the paper .II Related work : highlight literature review of Information Technology in health care .III Introduction of Diseases: which describe the disease its cause , symptoms and prevention and control for these (Hepatitis, Dengue, Acute respiratory, Kala-azar, Encephalitis) diseases. IV Process and Methodology: describes the process used, source of data taken ,tool used and techniques applied. V Conclusion: final results of analysis, VI Limitations and Future Scope: what are the limitation and challenges faced and further extension of study and its scope. Last section is for the reference.

### II. RELATED WORK

Prescod and James [1] worked on Neonatal Intensive care unit to study the mortality rate and then proposed a value for the same with respect to certain parameters which depict the health of new born child . The authors used AI technique to identify the same. They introduced this value or tool into NICU to improve health care delivery and mortality or survival predictions, which benefits parents also to understand the level of health of their new born child by answering a question “will my baby survive?”

Thakkar ,Hasan and Desai [2] were inspired to do this research by the study of mortality rate of swine flue. The paper focus on the aspect of medical diagnosis by learning predictions through the data collected for Swine flue. The authors proposed a method to identify swine flu by study 110 symptoms to decrease the cost incurred in the test of the disease .The author developed a prototype intelligent swine flue prediction software , they used Naïve Bayes classifier technique for classifying the patients of swine flue. Based on the possibility of the diseases and guaranteed the accuracy of almost 63.3 percent.

Chatterjee and Dutta[4] focus on Diabetes as a chronic disease, which if undetected may leads to acute and long-term complications and ultimately death. The author introduce a concept of “persuasive sensing” and the report is generated on the basis of mined data collected through a wireless sensor network system with in the home environment to capture their daily activities then they provide feedback via SMS/text (daily) and tailored newsletter (weekly). Author show that with the help of an Artificial Neural Network, we can predict blood glucose levels for the next day from accumulated data with an accuracy of 94%. The predictive model presented here is a break-through in at-home sensing research

Lu and Tong[5] proposed a framework on clustering analysis to explore health diseases topics from online health community automatically instead of using statistical analysis. Diabetes as one of the most common chronic diseases is discussed on this framework by using EM Clustering algorithm.

N. Wickramasinghe, S. Koritala [6] proposed a health care system as service provider for the better health of people based on the current health problems of American patients and their background.

### **III. INTRODUCTION OF DISEASES**

#### **A. HEPATITIS**

The word **hepatitis** comes from the Ancient Greek word *hepar* (root word *hepat*) meaning 'liver', and the Latin *itis* meaning inflammation. Hepatitis means injury to the liver with inflammation of the liver cells.

Hepatitis is of five types, A, B, C, D and E-plus types X and G.

Out of these A, C, D and E –plus types X and G are majorly caused by infected food or water whereas B is sexually transmitted disease caused by virus HBV.

#### **[1] SYMPTOMS:**

Hepatitis initial symptoms include a mild flu, and may include Diarrhea, Fatigue, Loss of appetite, Mild fever, Muscle or joint aches, Nausea, Slight abdominal pain, Vomiting etc

#### **[2] PREVENTION AND CONTROL:**

Wash your hands with soap after going to the toilet, consume cooked food, drink commercially bottled water, or boiled water, eat fruits that you can peel if you are somewhere where sanitation is unreliable, only eat raw vegetables if they have been cleaned/disinfected thoroughly Do not share needles, toothbrushes, Safe sex. Get a vaccine for Hepatitis.

#### **B. KALA-AZAR**

Word consists of “Kala” (in Hindi means “Black”) & “azar” (in Hindi means “Fever”) .Fever associated with dark complexion is also known as “Black Sickness” . It indicates terrifying effect of the disease on the imagination of the people rather than the actual reality of disorder.

It is most commonly found in tropical and temperate countries like India. Kala-azar is found in three forms as follows:

- Cutaneous (skin), the most common form
- Mucocutaneous (mucous membranes)
- Visceral (internal organs), a very rare form.

Kala-azar, is majorly caused by parasite which spreads through the bite of an infected female phlebotomine sand fly. The sand fly pick up the parasite Kala-azar from many common reservoir hosts such as domestic dogs, rats and humans. It usually feeds on an uninfected human, either at dusk or during the night. Certain species of parasites may spread through blood transfusions or contaminated needles if shared. It can even be Spread from a pregnant woman to her baby .

#### **[1] SYMPTOMS:**

Commonly, person suffering from Kala-azar receive nodules or sores on the skin and face. These symptoms may begin weeks or months after the bite of an infected sand fly. At the site of the bite, a papule is formed which eventually ulcerates. A crater appears at the centre of the ulcer and the edges are raised almost giving the ulcer, a volcano like resemblance. These lesions may occur in multiple. It may take months or years to heal. The associated lymph nodes in that area may also swell and lead to inflammation. It may also leave a residual scar.

#### **[2] PREVENTION AND CONTROL**

A combination of intervention strategies are needed to prevent and control Kala-azar. Key strategies include: Early diagnosis and effective case management, Vector control, Effective disease surveillance, Control of reservoir hosts, Social mobilization and strengthening partnerships,

#### **C ACUTE RESPIRATORY INFECTIONS (ARIS)**

Acute respiratory infections (ARIs) are classified as upper respiratory tract infections (URIs) or lower respiratory tract infections (LRIs). The upper respiratory tract consists of the airways from the nostrils to the vocal cords in the larynx, including the paranasal sinuses and the middle ear. The lower respiratory tract covers the continuation of the airways from the trachea and bronchi to the bronchioles and the alveoli.

#### **[1] SYMPTOMS**

ARIs are not only confined to the respiratory, tract but it may also have systemic effects due to extension of infection or microbial toxins, inflammation, and reduced lung function. Diphtheria, pertussis (whooping cough), and measles are vaccine-preventable diseases that may have a respiratory tract component but also affect other systems. Except during the neonatal period, ARIs are the most common causes of both illness and mortality in children under five, who average three to six episodes of ARIs annually regardless of where they live or what their economic situation is. However, the proportion of mild to severe disease varies between high- and low-income countries, and because of differences in specific etiologies and risk factors, the severity of LRIs in children under five is worse in developing countries, resulting in a higher case-fatality rate. Although medical care can to some extent mitigate both severity and fatality, many severe LRIs do not respond to therapy, largely because of the lack of highly effective antiviral drugs. Some 10.8 million children die each year. It is caused majorly in those areas where medical care is not available or is not sought

## [2] PREVENTION AND CONTROL

Interventions to control ARIs can be divided into four basic categories: immunization against specific pathogens, early diagnosis and treatment of disease, improvements in nutrition, and safer environments. The first two fall within the purview of the health system, whereas the last two fall under public health and require multisectoral involvement in Vaccinations, Case Management,

### D ENCEPHALITIS

**Encephalitis** (from Ancient Greek ἐγκέφαλος, *enképhalos* “brain”, composed of ἐν, *en*, “in” and κεφαλή, *kephalé*, “head”, and the medical suffix *-itis* “inflammation”) is an acute inflammation of the brain. Encephalitis with meningitis is known as meningoencephalitis..

#### 1. SYMPTOMS

Encephalitis is usually caused by Viral, Bacterial, Limbic system encephalitis, Autoimmune encephalitis, Encephalitis lethargica. Person suffering with encephalitis shows acute onset of fever, headache, confusion, and sometimes seizures. Younger children or infants may show symptoms like irritability, poor appetite and fever. Patient with meningitis can also get Stiff neck, due to the irritation of the meninges covering the brain.

#### 2. PREVENTION AND CONTROL

Treatment generally depends on the symptoms. Very few reliably tested specific antiviral agents are there (e.g. acyclovir for herpes simplex virus) and are used with limited success in treatment of viral infection, with the exception of herpes simplex encephalitis. In patients with high risk, sometime supportive treatment, such as mechanical ventilation, is required. Corticosteroids (e.g., methylprednisolone) are used to reduce brain swelling and inflammation. Sedatives may be needed for irritability or restlessness. For *Mycoplasma* infection, parenteral tetracycline is given. Encephalitis due to *Toxoplasma* is treated by giving a combination of pyrimethamine and sulphadimidine.

Vaccines against tick-borne and Japanese encephalitis are available and should be given to at-risk individuals.

### E . DENGUE

(UK /'dɛŋɡeɪ/ or US /'dɛŋɡi:/), also known as **breakbone fever**, is a mosquito-borne tropical disease caused by the dengue virus.

#### 1. SYMPTOMS

Generally, 80% infected people will only have mild symptoms such as an uncomplicated fever, 5% may have more severe illness and in a small proportion it is life-threatening.

Caused by virus (DENV) is an RNA virus of the family *Flaviviridae*; genus *Flavivirus*.

Dengue virus is primarily transmitted by *Aedes* mosquitoes, particularly *A. aegypti*. These mosquitoes usually live between the latitudes of 35° North and 35° South below an elevation of 1,000 metres (3,300 ft).

The period between exposure and onset of symptoms ranges from 3–14 days, but most often it is 4–7 days. Therefore, travelers returning from affected areas are unlikely to have dengue if fever or other symptoms start more than 14 days after arriving home. Children often experience symptoms similar to those of the common cold and gastroenteritis and have a greater risk of severe complications, though initial symptoms are generally mild but include high fever.

#### 2. PREVENTION AND CONTROL

There are no vaccines for the dengue virus. Prevention thus depends on control of and protection from the bites of the mosquito that transmits it. The World Health Organization recommends an Integrated Vector Control program consisting of five elements:

1. Advocacy, social mobilization and legislation to ensure that public health bodies and communities are strengthened;
2. Collaboration between the health and other sectors (public and private);
3. An integrated approach to disease control to maximize use of resources;
4. Evidence-based decision making to ensure any interventions are targeted appropriately; and
5. Capacity-building to ensure an adequate response to the local situation.

The primary method of controlling *A. aegypti* is by eliminating its habitats. This is done by getting rid of open sources of water, or if this is not possible, by adding insecticides or biological control agents to these areas.

## IV. PROCESS AND METHODOLOGY

**A .Data:-**Data is taken from INDIA STAT. fifteen tables were used to analyse the data of five diseases named as 'Hepatitis', 'Dengue', 'Acute respiratory', 'Kala-azar', 'Encephalitis' on twenty nine states and union territories of India then rate of mortality was calculated .

**B. Tools and techniques:-**Various types of data mining tools are currently available and each one has its own merits and demerits. For the analysis of the first objective we used clustering (K-mean Technique ) to find out the similar pattern of diseases in the cluster. This is used on 3 clusters and seed is chosen randomly. Second objective was to find out the relation between these 5 diseases we have used classification (apriori algorithm) with a rule based on 37% support and 92% confidence.

Then the inference is drawn as a conclusion and briefly represented under the section of conclusion of this paper. We used the TNAGRA tool for the same.

## V. CONCLUSION

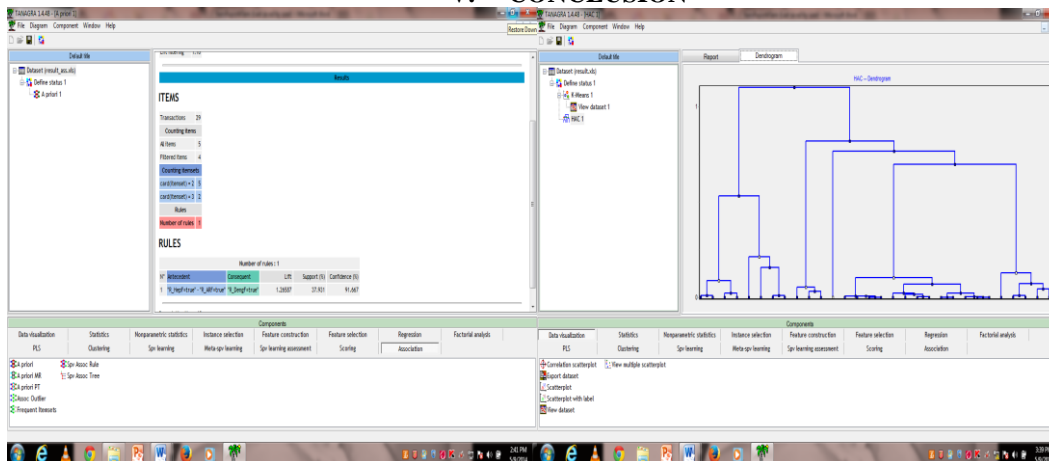


Fig1 : Result of Apriori Algorithm

Fig2:Result of K-mean Technique

- It has been concluded that if there is the existence of Hepatitis and Kala-azar then there must be high chances of Dengue.
- Based on the pattern of existence of diseases and rate of mortality in particular state we have categorised all states into 3 clusters and concluded that Himachal Nagaland and Sikkim are in same cluster , Jharkhand, Karnataka, Kerala and Madhya Pradesh , Maharashtra, Manipur and Mizoram are in same rest states are in different clusters Hence , the existence of rate of mortality in the states of each cluster is similar or nearby.

This analysis can be further extends to demographic or social and economic conditions of states

## VI. LIMITATIONS AND FUTURE SCOPE

### A. LIMITATIONS-

Healthcare data mining can be limited by the accessibility of data like raw data or ownership of data, Secondly, other data problems may arise which include missing, corrupted, inconsistent, or non-standardized data, such as pieces of information recorded in different formats in different data sources. The quality of data mining results and applications depends on the quality of data.

### B. FUTURE DIRECTIONS

Data mining applications in healthcare can have tremendous potential and usefulness. However, the success of healthcare data mining hinges on the availability of clean healthcare data. In this respect, it is critical that the healthcare industry consider how data can be better captured, stored, prepared, and mined. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications. Further, as healthcare data are not limited to just quantitative data, such as physicians' notes or clinical records, it is necessary to also explore the use of text mining to expand the scope and nature of what healthcare data mining can currently do. In particular, it is useful to be able to integrate data and text mining.

Finally, the authors hope this paper can act as a decision maker in order to identify that what all states are there where we need to focus in order to identify the cause of diseases it could be demographic , social or economic conditions of that states which may uses this study.

### ACKNOWLEDGMENT

We wish to acknowledge the help provided by Prof. ARVINDER KAUR, USIT, GGSIPU, New Delhi(INDIA)

### REFERENCES

- [1] F. I Prescod and Andrew James, "A Knowledge Management Based Approach for Mortality Prediction in the Neonatal Intensive Care Unit,"IEEE, Developments in E-systems Engineering Expert,2011
- [2] B. A. Thakkar,M. I. Hasan,,M. A Desai, "Health care decision support system for swine flu prediction using Naïve bayes classifier", International Conference on Advances in Recent Technologies in Communication and Computing,,IEEE,2010.
- [3] J. Adetunji,"Trends in under-5 mortality rates and the HIV/Aids epidemic",child mortality,Bulletin of the world health organization, 2000.
- [4] S. Chatterjee, K.Dutta, H. Byun, A. Pottathil," Persuasive and Pervasive Sensing: a New Frontier to Monitor, Track and Assist Older Adults Suffering from Type-2 Diabetes", 46th Hawaii International Conference on System Sciences,IEEE,2013.
- [5] Y. Lu,S. J. Tong,P.Zhang,S. Deng "Exploring Health-related Topics in Online Health Community using Cluster Analysis", 46th Hawaii International Conference on System Sciences,IEEE,2012.
- [6] N. Wickramasinghe,S. Chalasani,S. Koritala,"The role of health care system of systems and collaborative Technologies in providing superior healthcare delivery to native American Patients", 45th Hawaii International Conference on System Sciences,IEEE,2012.