DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5261

ISSN No. 0976-5697

Volume 9, No. 1, January-February 2018

International Journal of Advanced Research in Computer Science

SURVEY REPORT

Available Online at www.ijarcs.info

SURVEY ON MINING DIABETES DATA AND ITS APPLICATIONS ON DIAGNOSING METHODS IN DISEASE MANAGEMENT USING BIG DATA

Puneeth Thotad
Asst. Prof. & Research Scholar, Dept. of Master of Computer Applications, KLEIT, Hubballi, India

Dr. Shashikumar G Totad,
Professor, Department of Computer Science and Engineering, BVB, Hubballi, India

Dr. Geeta R. Bharamagoudar
Professor & Research Advisor, Dept. of Information Science and Engineering, KLEIT, Hubballi, India

Shanta Kallur
[4]Asst. Professor, Department of Computer Science and Engineering, KLEIT, Hubballi, India

Abstract : Diabetes is one of the Noncommunicable Disease (NCD), and the major health hazard in countries like India, China etc. The serious problem associated with DM is also associated with short-term symptoms like Hypoglycemia, Ketoacidosis and long-term macro and micro vascular complications including cardiovascular diseases like Heart attacks, Strokes, Kidney failure, Diabetic foot diseases leading to gangrene, amputations and Blindness. Doctors face difficulties to predict future health conditions of patients' without relevant information. Thus need for Technology leveraging in Diabetes management is very essential. As the diabetes data collected from Health Industry or directly interacting with the patient is unstructured in nature. To analyses and prognosis it is necessary to organize the data and emphasize its size into nominal values. The effectiveness of Big Data analytics and Data Mining techniques in envisaging the risk of next occurring event in diabetic patients and taking proper precautions to overcome the risks is prophesied. The aim of this survey is to recall the progresses carried out in the area of disease management using Big Data technology.

Keywords: Diabetes is one of the Noncommunicable Disease (NCD), and the major health hazard in countries like India, China etc

I. INTRODUCTION

Electronic Health Records (EHR) is made up of Patients' data in the digital format consisting of valuable information but heterogeneous and inconsistent data. Analytics can be applied to identify bottlenecks and enhance the patient treatment efficacy. With the help of technological developments, it is necessary to combine robust diabetic data sharing. healthcare gadgets can be used to access health services at all stages of patient conditions. To design the predictive model, it requires working with unstructured, large volume of data. In today's arena of Electronic Health Informatics, Big Data implementation can help us to increase the focus on understanding and development of predictive tools to manipulate and analyze data sets to correlate and collate insights to facilitate better understanding of the issue that lead to next event occurrence and take initiatives to control the severity of disease. Constructing predictive modeling solution is highly challenging for recognizing risk involved in next occurring event.

Big Data is a new revolutionary idea, which deals with large set of voluminous, unstructured, real-time data in Tera bytes.

Big Data analytics in Health Care involves large volume of Data collection, analyzing and providing various ways to retrieve relevant information regarding patients' health conditions. Predictive analytics helps to take more proactive approaches for treatment. It is possible to predict which patient will develop chronic conditions. It helps to analyze which age groups of people are much affected by DM and its consequences. The medical examiners can focus not only on treating existing conditions but also, preventing recurrences. Hospitals can be well prepared to ensure that there are enough resources in hand to provide utmost care. It also improves the outcomes of patient's treatments. By closely analyzing which treatment works best, providers can predict the next occurring event and make more intelligent decisions about treatment. Here complications involved in treatment can be minimized [1].

Table 1: The IDF report of Diabetes in India for the year 2015

Total Adult Population (1000s)	798,988	Number of deaths in adults due to diabetes	1,027,911
Prevalence of diabetes in adults (20- 79 years) (%)	8.7	Cost per person with diabetes (USD)	94.9
Total cases of adults (20- 79 years) with diabetes (1000s)	69,188.6	Number of cases of diabetes in adults that are undiagnosed (1000s)	36,061.1

It is estimated that 371 million people worldwide have diabetes and the number is increasing at an alarming rate. It is observed that prevalence of diabetes increases tenfold, from 1.2% to 12.1% between 1971 to 2000. According to the International Diabetes Federation's (IDF) survey 61.3 million people aged 20 to 79 years live with diabetes in India by year 2011 and this number is expected to increase to 101.2 million by 2030 [2]. According to IDF's survey in 2015, 415 million

people have diabetes in the world and 78 million people in the South East Asia (SEA) Region; by 2040 this will rise to 140 million. There were 69.1 million cases of diabetes in India in 2015 and the details are as shown in Table 1.

II. NEED FOR TECHNOLOGY LEVERAGING IN DIABETES MANAGEMENT

Diabetes is a metabolic disorder characterized by chronic hyperglycemia with disturbances of carbohydrates, fat and protein metabolism. It is found in persons of all age group. It varies from person to person depending on the symptoms and levels of blood sugar in human body. The serious problem of Diabetes is associated with short-term symptoms like Hypoglycemia, Ketoacidos is and long-term macro and micro vascular complications including cardiovascular diseases like Heart attacks, Strokes, Kidney failure, Diabetic foot diseases leading to gangrene and amputations and Blindness.

The reasons for diabetes are mainly due to changing lifestyle of people, it can lead to decreased physical activity. Improper diet with consumption of fats, high calories of sugar items, and higher stress levels affects insulin sensitivity and obesity. Very high costs is incurred in diabetes treatment. According to World Health Organization (WHO), if the adult in the low income family has Diabetes, 25% of family income may be devoted to diabetic care alone. Diabetes is classified it to Type 1, Type 2 and Gestational Diabetes [3].

Type 1 Diabetes Mellitus: It is also called Juvenile diabetes. Type 1 diabetes typically occurs in young adults and children. A chronic condition in which the pancreas produces little or no insulin or it happens when your immune system destroys Beta cells in pancreas. It typically appears in adolescence. Symptoms are increased thirst, fatigue, hunger, frequent urination, blurred vision and weight loss.

Type 2 Diabetes Mellitus: Type 2 diabetes develops most often in middle-aged and older people who are also overweight or obese. It is a long term metabolic disorder that is characterized by high blood sugar, insulin resistance and the way how the body handles glucose. Symptoms include Polyuria, weight loss, Yeast Infection and Blurred Vision. The disease, once rare in youth, is becoming more common in overweight and obese children and adolescents. Scientists think genetic susceptibility and environmental factors are the most likely triggers of type 2 diabetes.

Gestational Diabetes Mellitus: Gestational diabetes is a condition in which a woman without diabetes develops high blood sugar levels during pregnancy. **Gestational diabetes** generally results in few symptoms; however, it does increase the risk of pre-eclampsia, depression, and requiring a Caesarean section. Women with unmanaged Gestational Diabetes are at increased risk of developing Type 2 Diabetes.

II. I Complications of Diabetes Mellitus

If patient doesn't make an effort to get control on Diabetes, It leads to many complications. Diabetes can take a toll on nearly every organ in your body, including the:

- **Heart related problems**: Heart disease and blood vessel disease are common problems for many people who don't have their diabetes under control. Person with diabetes might have at least twice as likely to have heart problems and strokes [4].
- **Blood vessel damage or Nervous system damage** : It may cause foot problems that leads to amputations. People with diabetes are more likelier to have their feet and legs removed.
- **Eyes related problems:** Diabetes is the leading cause of new vision loss among adults ages 20 to 79. It can lead to eye problems, some of which can cause blindness if not treated: Glaucoma, Cataracts, Diabetic retinopathy [5].
- **Kidneys related problems:** Diabetes may also cause of kidney failure in adults. Patients might not notice any problems related to kidney in the early diabetes stage. In later stages it can make your legs and feet swell.

Caring for patients with diabetes, or any chronic disease, particularly newly diagnosed, can be time-consuming for Physicians, but technology can be leveraged to help them provide care in an efficient and organized manner.

Data Required in Health Care Analytics:

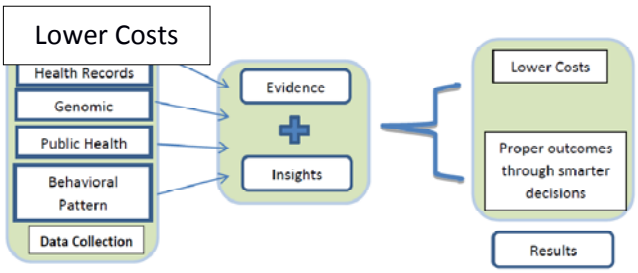

Figure 1: Big Data implementation in Health Care

III. BIG DATA IN HEALTH CARE INDUSTRY

Big Data equip patient centric services to deliver faster relief to the patients by providing evidence based medicine. It helps to detect diseases at the earlier stage based on the clinical data available, Minimizing the drug doses to avoid side effect and provide efficient medicine based on genetic Makeups. This helps in reducing readmission rates and avoids future consequences in health condition of patients there by reducing the cost for the patients.

Improving the treatment methods by providing the customized patient treatment by monitoring the effect of medication continuously and based on the analysis dosages of medications can be changed for faster relief. Analysis of the data generated by the patients already suffered from the same symptoms helps physicians to provide effective medicines to patients [6]. The data can be obtained from interacting with patients, clinical system, pharmaceutical records, Physician notes, Laboratory reports, X- Ray reports, case history, medical images, diet regime, sensor data, social logs etc. as shown in Figure 1.

III.I Hadoop

Hadoop isan distributed data processing platform. Hadoop can be applied on large set of variable size data to organize and perform analyses. It can process large amount of health data by allocating each block of data sets to large number of servers which act like clusters, used to salve the lock in the large problem and then integrates to obtain the final result. The modern technologies like Hive, Zoo Kipper and Pig can be implemented on the Hadoop Distributed File System

(HDFS) to compute more Knowledgeable information. Hadoop has two main components to use for computation of this job: Map Reduce and HDFS.

- **Map / Reduce**: Map/ Reduce principle is based on programming models to process diabetes data by dividing them into small blocks of tasks. It uses distributed algorithms, on a group of computers arranged in a cluster, to process large datasets. It contain two operations:
  - Map(): This function being on master node, divides the input data or task into smaller subtasks, which it then distributes to slave nodes that process the smaller tasks and pass the results back to the master node. The subtasks run in parallel on multiple computers which act like slave nodes.
  - Reduce (): This function collects the results of all the subtasks obtained from the slave nodes and combines them to produce an aggregated final result, which in returns as an answer to the original big query.

- **HDFS Architecture:** Hadoop effectively handles the large data set. The name node communicates to Job Tracker and assigns the task given by the client. Map Reduce program performs the analysis on the data and returns the results to job tracker. It also returns the block where the client can store its data.

- **Name node:** It is the master node which receives the request from the patient monitoring system (client). It looks up the Meta data to find out which is the suitable data node for storing the data related to the client. It selects data node based on the locality and available free slots.
- **Job Tracker:** Map reduce program running in job tracker assigns job to the data node and task tracker. Data node stores the actual data and it periodically sends heartbeat to name node about the data stored. Task tracker compute the task assigned by job tracker.

IV. CURRENT WORKS CARRIED IN PROGNOSIS AND DIAGNOSIS OF DIABETES MELLITUS USING DATA MINING AND BIG DATA

A literature review reveals many results on diabetes carried out by different methods and materials of diabetes problem in India. Many people have developed various prediction models using data mining to predict diabetes. Prediction analytics is a key technique that will extend the clinical and treatment decisions in health care field. It is observed that one big challenge is to get access to a very large database which is scalable to petabytes, to accommodate patients' data from multiple sources [1].

Combination of genetics, classification, regression, and neural network is used for handling the missing and out lier values in diabetic data set. Also, they have replaced the missing values with domain of the corresponding attributes. The neural network model is used for prediction, on the preprocessed dataset [7]. In predictive analysis of diabetic treatment using regression based data mining techniques discover patterns using Support Vector Method algorithm, which identify the best mode of treatment for diabetes across different age groups [8]. In conclusion the drug treatment for patients in the young age group can be delayed where as patients in the old age group should be prescribed drug treatment immediately. The emergency of medical record (in large scale) database presents opportunities for database personalized medicine. The sequence of medical conditions can be predicted in the patient based on the data collected form the patients [9]. Hierarchical association rule model (HARM) is an approach to predict the next event within the current sequences given a database of past sequences. It generates a set of association rules such as:

Dyspepsia + Epigastria pain ➔ Heart Burn

It says that in general dyspepsia and epigastric pains are followed by heart burn. HARM yields a prediction algorithm for sequential data can be used for several applications beyond condition prediction. Big data analytics have been applied to evaluate the risk of readmission of diabetic patients. Predictive modeling has been employed by applying decision tree classification method to improve the performance. The importance of each variable in diabetic dataset is identified and analyzed [10].

Big Data analytics provides systematic way for achieving the better outcomes like availability and affordability of health care services to all population [11].In Pima Indian Diebetis Database, the various types of diabetes predeicted and classified using C4.5 Classification algorithm[12]. Hive and R launguages were used efficiently to analyse Pima Indian diabeted dataset. It also helpedto develop some prediction models [13].

Using clinical data, Hypothesis generation tools and PubMed trends the association between diabetic retinopathy and antihypertensive drugs is discovered [5]. The comparison between the diabetic retinopathy and antihypertensive drugs, with the medication patterns in a clinical population revealed that though the literature is dominated by drugs developed more recently (ACE inhibitors and ARB's), older drugs are still widely prescribed in some clinical environment. In prediction of treatment effect in diabetic patients using different classification methods like Decision Tree, Rule induction and Naive Bayes model using Rapid Miner tool. It is proposed that providing patients with suitable environment to stay in are crucial can change the result of the treatment [14].

For analysis of diabetes data, researchers have worked and developed efficient prediction models. The predictive analysis involves the methods based on three areas such as Operations Management, Medical Management and biomedical, and System design and planning separately but it's not possible to find one method which includes all three.

So far experiments in Diabetes Mellitus are limited to predicting the risk of readmissions and treatment effect in diabetic patients. The possibility of using Data Mining and Big Data Paradigm in prognosis of patient health condition to take an adequate measure is the future research. In order to have control over patient health factors and reduce the cost incurred on patients a research topic titled "Mining Diabetes Data for Early Prognosis and Diagnosis in Disease

Management" is formulated. The following are the objectives of the proposed research topic.

Creating systems that can process subjective information effectively requires to overcome number of novel

V. Steps involved in Big Data implementation in Prognosis and diagnosis techniques in Diabetes Mellitus

The aggregation of individual patients Health information into large data sets allow usage of specialized software tools and applications for predictive analytics and data optimization. By analyzing large data sets, the meaningful and useful patterns can be derived to treat patients. The proposed research topic using Data Mining and Big Data analytics is a revolutionary technology for diabetic data analytics, which deals with large sets of unstructured real-time data in Terabytes. Using predictive data analytics in health care is highly appropriate as health case records are very large and unstructured datasets. The next occurring event using predictive analytics system includes data collection, data processing, Algorithms for data modeling and processing the analyzed reports [16].

The different stages of proposed system for Diabetes prognosis and Diagnosis using Data Mining and Big Data techniques are shown in Figure 2.

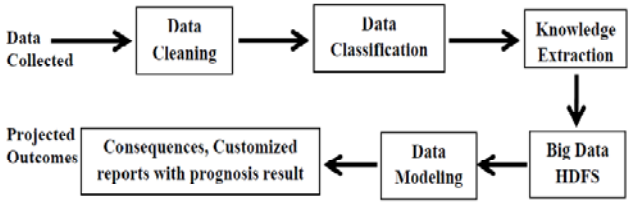

Figure 2: Framework of a typical model for Diagnosis of Diabetes and prognosis of secondary impediments.

➢ **Data Collection:** This process collects data from various sources as mentioned in Figure 1 and stores it in Hadoop Distributed File System (HDFS). The raw diabetic data set will be given as the input to the system obtained from various data repository or also, data can be obtained from interacting with patients, clinical system, pharmaceutical records, Physician notes, Laboratory reports, X- Ray reports, case history, medical images, diet regime, sensor data, social logs etc.

➢ **Data Cleaning:**Data cleaning reduces errors and improves the data quality. Such data need to be removed or some simulated data can be added to fill the gap. Data cleaning is a process used to determine inaccurate, incomplete or unreasonable data and then improve the quality through correcting of detected errors and omissions.

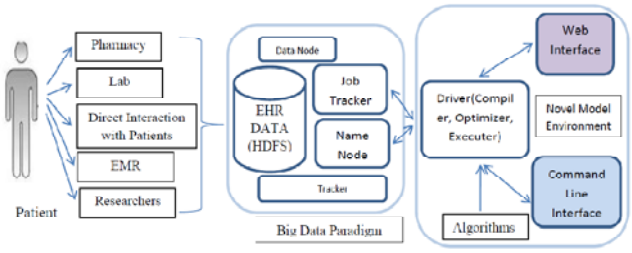

Figure 3: Big Data technology used in Health care systems.

➢ **Data Classification:** The diabetes data will be classified into Structured, Semi- structured and unstructured to perform meaningful analysis. The process might include Association rule mining, Clustering of data with similar patterns, classification of health risk value by level of patient health condition, and usage of previous static information to identify the next occurring event by passing to the Big Data environment as shown in Figure 3.

➢ **Knowledge Extraction:** Itis creation of knowledge from structured (relational databases, XML) and unstructured (text, documents, images, physician notes) sources. The resulting knowledge needs to be in a machine-readable and machine-interpretable format and must represent knowledge in a manner that facilitates inference.

➢ **Data Modeling:** In this phase, the performance of classified data is analyzed. It involves a progression from conceptual model to logical model to physical schema.

➢ **Projected Outcomes:** It helps in generation of reports based on the results obtained from the above data modelling techniques. Provides a web based user interface application, in which users/ patients' valid data can be entered and get the prognoses reports. This report can be further referred by physicians' to supervise the patients.

The evidence for a medication or disease is derived by aggregating thousands of patient's records to find the future occurring event in diabetic patients. The meaningful use of Data Mining and Big Data methodologies will help to realize the true results and improve the safety, quality health of diabetic patients.

VI. CONCLUSION

Technology leveraging in better diabetes management will be achieved. More importantly, physicians and patients are at the advantage of the proposed research. Personalized insulin and medicine dosage helps the patients to take more preventive measures and avoid further complications. Finally, developing a system around the research findings will be useful in prognosis and diagnosis of diabetes based on the patients' data and improve the quality of life lead by the human kind.

REFERENCES

[1] HarithaChennamsetty, Suresh Chalasani, Derek Riley. "Predictive Analytics on Electronic Health Records (EHRs) Using Hadoop and Hive" IEEE publications-978-1-4799-6085-9-2015.

[2] David R. Whiting, et al. "IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030", Diabetes Research and Clinical Practice, Volume 94, Issue 3, Pages 311-321, December- 2011.

[3] KonstantionsMakris, LoukiSpanou, "Is there a Relation between Mean Blood Glucose and Glycated emoglobin?", Journal of Diabetes Science and Technology, Volume-5, Issue 6, Nov-2011.

[4] Sujata Rani, Mrs. ShagunGirdhar, "Analysis of Heart Disease and Diabetes Using Artificial Neural Networks", International Jiurnal of Advanced Research in Computer Science and Software Engineering, Volume-4, Issue 8, Aug-2014.

[5] Katherine Senter, Sreenivas R. Sukumar, Robert M. Patton and Edward Chaum."Using Clinical Data Hypothesis Generation Tools and PubMed Trends to Discover the Association between Diabetic Retinopathy and Antihypertensive Drugs" IEEE International Conference on Big Data. PP. 2578-2582. 2015

[6] J. Archenna, E. A. Mary Anita. "A Survey of Big Data Analytics in Health care and Governament". International Symposium on Big Data and Cloud Computing- 2015, Pages 408-413, 2015

[7] V. H. Bhat, P. G. Rao, and P. D. Shenoy, "An Efficient Prediction Model for Diabetic Database Using Soft Computing Techniques," Architecture, Springer-Verlag Berlin Heidelberg, pp. 328-335, 2009.

[8] Abdullah A. Aljumah, Mohammed GulamAhamad, Mohammad Khubeb Siddiqui, "Application of data mining: Diabetes health care in young and old patients", Journal of King Saud University – Computer and Information Sciences, vol. 25, pp. 127–136, 2012.

[9] Tyler H, McCormik, Cynthia Rudin and David Madigan. "Bayesian Hierarchical Rule Modeling for predicting medical conditions" in the Annals of Applied statics', Volume 6, no 2, PP. 652-668, 2012.

[10] SaumyaSalian, Dr. G. Karisekaran. "BigData Analytics Predicting Risk of Redmissions of Diabetic Patients" in International Journal of Science and Research (IJSR), Volume 4, no 4,pp. 534-538, 2015.

[11] Dr. Saravana Kumar N. M., Eswari T, Sampath P. and Lavanya S. "Predictive Methodology for Diabetic Data Analysis in Big Data" in 2nd International Symposium on Big Data and Cloud Computing (ISBCC' 15) pp. 203-308, 2015.

[12] K. Rajesh, V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis" in International Journal of Engineering and Innovative Technology (IJEIT) Vol 2(3), 2012.

[13] Sadhana, Savitha Shetty, "Analysis of Diabetic Data Set Using Hive and R", International Journal of Emerging Technology and Advanced Engineering, Volume-4, 2014.

[14] HaotingXhang, Lingun Shao, "Predicting the treatment effect in diabetic patients using classification models" International Journal on Digital Content Technology and its Applications (JDTA), volume 8, Pages 136-143, Oct- 2014

[15] Anjana RM, Pradeepa R, Deepa M, Datta M, Sudha V, Unnikrishnan R, et al. "Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: phase I results of the Indian Council of Medical Research-India Diabetes (ICMR-INDIAB) study" Diabetologia 54.12 2011.

[16] M. Gowsalya, K. Krushitha, C. Valliyammai. "Predicting the Risk of Readmission of Diabetic Patients using MapReduce" International Conference on Advanced Computing (ICoAC)- IEEE- 2014.

[17] Raghunath N. Adhiraj S., et. al., "A look at Channlenges and Apportunities of BigData Analytics in Health Care", IEEE-International Conference on BigData-2013.