ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

ARACHNO – A Deep Inspection Predictive Analysis Algorithm for Analyzing Diabetic data

R. Vanathi[1], Dr. A. Shaik Abdul Khadir[2]

Research Scholar, P.G. and Research Department of CS, Khadir Mohideen College, Adirampattinam, Thanjavur (Dist), Tamil Nadu, India[1]

Associate Professor, P.G. and Research Department of CS, Khadir Mohideen College, Adirampattinam, Thanjavur (Dist), Tamil Nadu, India[2]

ABSTRACT: The healthcare industry is an ever growing field, generating trillions and trillions of data every day. The modernization of the sector has a direct connection to this incremental proportion. These obtained data sets are partially structured but mostly unstructured in nature. These obtained data has to be processed with utmost care to derive complete usable patterns for subjective and predictive studies. These massive records of data, after processing, when employed, will prove to be highly complex. Since the employed data sets are unstructured in nature, the analysis of such data sets must be constrained to two major factors namely nominal size and emphatic miniaturization of data sets which will give them a structural aspect. Healthcare industry faces numerous hurdles that make us necessitate the discovery of sustainable data analytic method. Diabetes mellitus (or diabetes) is a chronic condition that deteriorates the body's ability to absorb the energy present in the food. There are three major types of diabetes: type 1 diabetes, type 2 diabetes, and gestational diabetes. The severe complexion of Diabetes mellitus is concomitant with long term problems. In this paper, we propose ARACHNO, a deep inspection predictive analysis algorithm. This algorithm is presented in Hadoop/MapReduce ecosystem. The proposed algorithm will predict the diabetes mellitus types that are most common, the problems concomitant with that type and the treatment that can be recommended. The output of this analysis is a deep inspection system that provides accurate treatments which is effective and also efficient.

KEYWORDS: Healthcare sector, Hadoop, MapReduce, Big data Analytic, Electronic Health Records, Data Prediction, algorithm.

I. INTRODUCTION

With the advancement in the field of medical science, the death rate has been decreased to a maximum amount. This decrease is directly proportional to the innovative and efficient discoveries that have been made in this field. The healthcare industry has generated large amounts of data. This humongous data generation is attributed to the growth in patient records on a day to day basis. The generated data encompasses both patient and their respective health practices that have been prescribed to them. Most of the time, it is the small quantity of medical records that serves as the basis of the high proportion of health care practice methodologies. Authentic tactical identification may expedite tailored arbitration [1]. Health records may be of any form and it may be in structured or unstructured format. It may be Electronic Health Record (EHR), Analytical Test Report (ATR), Imaging Records or Patient's Medical Journals. These records must be consistent and must be secure at all cost. In recent times, with the advent of numerous technologies like Internet of Things and Wearable sensors, the medical treatment procedures have reached a whole new level. A patient's medical sensor data can be updated any time and can be monitored in real time. These data sets can also be uploaded to the concerned medical servers and can be facilitated to centralized sharing. The important factor that is looked upon by

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

the patient's is the concealed acquaintance of their medical records [2]. These medical advancements have also made the treatment providers to create a medical database portal for their patients.

With the help of these portals, the patients can access their respective medical records and can also update any progress/egress in their recovering ability. This system empowers the patients with the ready form information availability to ensure concise treatment procedures and also enables the health care adviser to prescribe exact treatment based on their current status [3]. Such advancements in the field of medicine have lead to a new field of treatment known as Personalized medicine. In the field of Personalized medicine, a patient's medical record is analyzed deeply and thoroughly for sequencing the therapeutic medical procedures and then these procedures are applied in a monitored and analytical way. A clear study is made after the application of these procedures and the obtained results are analyzed for deviations [4].

Big data analytics has improvised these practical medical procedures and have alleviated the need of constant monitoring of the patient's health records as the entire medical record monitoring systems are automated. These analytics can be directed in two ways namely medicinal procedural ways or in historical treatment record diagnosis ways. Sometimes meager importance is given to analyzing the business value acquaintance of the data that is processed to be analyzed. Healthcare practitioners can get new intuitions and forecasts based on these records and can deliver better treatment to the patients [5]. Big Data Analytics is greatly driven by its characteristics which defines the adeptness and consistency of the analyzed data. These characteristics include Volume, Variety, Velocity and Veracity. With the constant rise in the healthcare industry data sets, these characteristics teem up to be of greater importance as the source of data generation is humongous and requires constant analyzing for discovering unknown patterns. When such abundant quantity of data is analyzed, we can extract much more useful patterns which can be applied for the betterment of the medicinal procedures. When such advancements join hands with high technologies, the attained level can only be maintained with constant check for deviations. As the health care industry advances, humongous amount of data is generated as Electronic Health Records (EHR). These EHR's must be carefully organized and studied to enhance the quality of the medical practices [6]. When such studies of patterns are made, the research must be handled by knowledgeable healthcare sector professionals so that the design, development, enrichment and advancement of the research will be state of the art. It also means that the patients care and patients future diagnostic methodologies will reach a new dimension [7]. This evolution of healthcare medicinal practices will directly involve the change in the field of drug discovery and pharmacology. It also acts as a value adder for the field of Pharmacogenetics which implicates the effect of genetic factors with relation the drugs taken and epidemiology which deals with the occurrence, administration, and possible governing of diseases and other antecedents relating to health.

The upsurge in magnitude and diversity of data that can be obtained in collaboration with the amendment of packing capacities and tools of analytical examination tender plentiful prospects to all the involved stakeholders which includes medicine manufacturers, medicinal practice regulators, buyers, Institutional and Individual healthcare providers, strategic decision formulators, researchers but most significantly, it has the capacity to enhance general health consequences if we acquire the knowledge of how to handle and manipulate it in a prioritized way[8].

Diabetes is a prolonged disease that occurs either when the pancreas does not secrete enough insulin or when the body is not in a position to put the secreted insulin in a proper use. Insulin is a hormone that does the blood sugar regulatory mechanism. There can be two different medical conditions of diabetes namely Hyperglycemia also known as High blood sugar or Hypoglycemia also known as low blood sugar. On a Global scale, an estimated 422 million adults were living with diabetes in 2014, compared to 108 million in 1980. The global occurrence (age-standardized) of diabetes has nearly made a rapid growth since 1980, rising from 4.7% to 8.5% in the adult population. This is directly proportional to the increase in associated risk factors such as being overweight or obese. Over the past decade, diabetes prevalence has risen faster in low- and middle-income countries than in high-income countries. Diabetes caused 1.5 million deaths in 2012. Higher-than-optimal blood glucose caused an additional 2.2 million deaths, by increasing the risks of cardiovascular and other diseases. Forty-three percent of these 3.7 million deaths occur before the age of 70 years.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

II. RELATED WORK

A deep study on various research works related to diabetes studies and diabetes prediction models reveals a wide array of research methodologies that are involved in the diabetes research examination process.

Nasim Samandari et al propose a research work which aims to identify circulating microRNA (miRNA) that predicts clinical progression in a cohort of 123 children with new-onset type 1 diabetes mellitus. For this research work, the research team collected plasma samples from 40 children who were diagnosed from a subset of from the Danish Remission Phase Cohort, and profiled for miRNAs. The plasma samples were taken at 1, 3, 6, 12 and 60 months after diagnosis. The clinical procedure they carried out to predict the nature of study is a viable one and the sequential study paves way for a much broader future works [11].

Research team comprising of Danqing Zhao, Liming Shen, Yan Wei, Jiaming Xie ,Shuqiang Chen, Yi Liang, Youjiao Chen and Haorong Wu made a significant study to identify serum biomarkers, which can predict the subsequent development of Gestational diabetes mellitus (GDM) at early stages. The experimental design was to collect blood samples from pregnant women at 12 to 16 weeks of pregnancy and perform iTRAQ analysis. In the results, Thirty three differentially expressed proteins were identified between case and control groups [12].

Yan-Hui Li et al proposed a machine-learning algorithm to predict new Hypertension genes as Hypertension is a major cardiovascular risk factor and accounts for a large part of cardiovascular mortality. Hypertension genes enrich in certain biological processes and show certain phenotypes. The proposes algorithm is concise and can be referenced to any work related to the research of hypertension gene manipulation [13].

Researchers Phattharat Songthung and Kunwadee Sripanidkulchai mined the data-sets collected from various healthcare practitioners to find the individuals having high risk potential to the onset of type 2 diabetes. They used Rapid Miner Studio 7.0 with Naive Bayes and CHAID (Chi-squared Automatic Interaction Detector) Decision Tree classifiers and was able to predict the high risk individuals. This prediction was later compared to hand-computed diabetes risk scoring mechanisms. Their method of using classification yielded better results than hand computed mechanisms [14].

Tanvi Anand et al analysed the helpfulness of employing techniques used in data mining to health care informatics. They also made an evaluation of varied approaches and techniques used in healthcare sector. The main motive of this research is to manipulate a machine driven tool for recognizing and distribution of important and appropriate healthcare information [15].

Peter Groves and his team of efficient researchers made a study on the pellucidness made by the healthcare sector major players on the healthcare industry data-sets. This accomplishment has lead to a state of data getting to the state of convertible assets. The research paves a new scope of dimension where all the healthcare industry data-sets are made transparent and the improvements in the technologies tend to exploit these data-sets into applicable and profitable knowledge [16].

Sarah DuBrava et al made a research on identification of volatiles concerned with the diagnosis of Diabetic Peripheral Neuropathy (DPN). This research is done by doing retrospective analysis on the samples that are chosen based on random forest modeling approach on the give EHR's. Accuracy of the proposed research model is evaluated using Receiver Operating Characteristic curves (ROC).This research gave a lucid way of the usage of retrospective analysis for determining the likelihood of a DPN diagnosis [17].

Researchers Priyanka Shetty and Sujata Joshi proposed a diabetes prediction and monitoring system using ID3 classification algorithm. In this prediction system, the research work is done by giving the diabetes symptoms as an input to the prediction model and the prediction is made. The monitoring part is powered by inputting the patient's medical reports and proper knowledge about the current status of the disease is presented to the patient as an easily understandable report [18].

Sajida Perveen et al proposed a research work follows adaboost and bagging ensemble techniques using J48 (c4.5) decision tree as a base learner along with standalone data mining technique J48 to classify patients with diabetes mellitus using diabetes risk factors. This classification is done across three different ordinal adults groups in Canadian Primary Care Sentinel Surveillance network [19].

Sometimes the clinical data can also be collected via cloud based applications. For example, Nadi Aridhal,a programmatic cloud suitable medical diagnostic computing application with optimal pulse system measurement algorithm which actually gets a pulse modified into 32 bit form data via sensors and plot it as time varying graph.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

All the above researchers have been fortuitous in examining the diabetic data set and unfolding new prediction models. In this paper, we use a deep inspection predictive analysis algorithm in Hadoop/Map Reduce ecosystem to predict and classify the type of diabetes. This system provides an efficient way to heal and tend the patients with better aftereffects at an affordable and viable way.

III. ARACHNO - A DEEP INSPECTION PREDICTIVE ANALYSIS ALGORITHM

We propose ARACHNO [Analytical Reporting and Auto-responding Cache Hard-token Network Object] algorithm for prediction and classification of diabetes. The proposed algorithm is made to run in the predictive analysis part of the proposed architecture.

A. Proposed Architecture

The proposed architecture employs Hadoop/MapReduce ecosystem. The proposed ARACHNO [Analytical Reporting and Auto-responding Cache Hard-token Network Object] algorithm works at the third phase of the architecture and is presented in a Hadoop/MapReduce ecosystem. The detailed architecture is presented in the Figure 2.

(i) Hadoop

Hadoop is an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license. The main objective of Hadoop is to facilitate storage and processing of large amount of data [21]. The method used by Hadoop is very unique. It uses a fission-fusion methodology as most of the time the analytical data that is fed in for prediction is of unstructured nature. Many applications of Hadoop nature has been created and the implication use of these applications are numerous. In the recent decades, there are numerous complementary applications developed by the open source community. These applications are both job oriented and also task oriented in nature with high importance being given to the hardware employed in the operational procedure [22].

(ii) MapReduce

MapReduce provides automatic parallel processing and distributed input output scheduling. MapReduce also facilitate Load balancing functions for the distributed input output scheduling. It also provides data transfer optimization in case of multiple databases employment and Network optimization in case of remotely distributed data processing. The main aspect of MapReduce is that it supports fault tolerance [23]. Apache Hadoop is the open source implementation of MapReduce. The following figure (Figure 1) gives a clear illustration of MapReduce workflow.

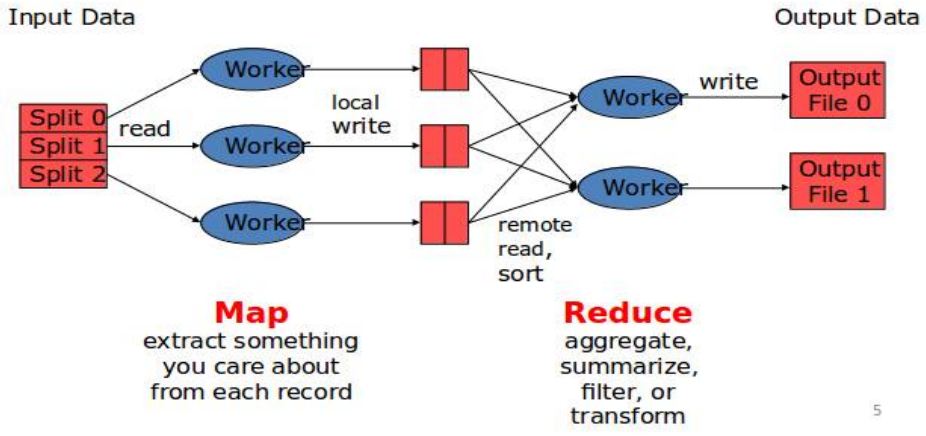

Figure 1: MapReduce Workflow

B. Data Collection

The raw diabetic data sets are given as input to the system. The diabetic patient's data can be obtained from UCI machine learning repository [25]. The data obtained from this repository comprises of both structured and unstructured data sets.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

C. Data set Characteristics

These data-sets comprises of data's from both EHR's and also patient's paper records. These records consists of four fields per record namely date of record created, time of record creation, code indicating the case of medical report and finally the value[25]. The code field can be interpreted as given in table 1.

Code	Interpretation
33	Regular insulin dose
34	NPH insulin dose
35	UltraLente insulin dose
48	Unspecified blood glucose measurement
57	Unspecified blood glucose measurement
58	Pre-breakfast blood glucose measurement
59	Post-breakfast blood glucose measurement
60	Pre-lunch blood glucose measurement
61	Post-lunch blood glucose measurement
62	Pre-supper blood glucose measurement
63	Post-supper blood glucose measurement
64	Pre-snack blood glucose measurement
65	Hypoglycemic symptoms
66	Typical meal ingestion
67	More-than-usual meal ingestion
68	Less-than-usual meal ingestion
69	Typical exercise activity
70	More-than-usual exercise activity
71	Less-than-usual exercise activity
72	Unspecified special event

Table 1: Dataset codes and its interpretation

D. Data Warehousing

In this phase the data-sets are assembled into a single sector. Here the data is treated and cleansed with all the required attributes. EHR's are processed in a correlation with their respective distributional EHR so that the unidentified patterns can be obtained. These processed data-sets are then seeded and delivered into the prediction system for the obtainment of the results.

E. Predictive Analysis

In this phase the data-sets that are forwarded from the Data warehousing phase is accumulated and are subjected to the ARACHNO [Analytical Reporting and Auto-responding Cache Hard-token Network Object] algorithm for prediction and classification of diabetes. Figure 2 illustrates the ARACHNO [Analytical Reporting and Auto-responding Cache Hard-token Network Object] algorithm. The output of this algorithm is detailed to the extent where not only the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

diabetes type is identified but also the medication prescriptions along with treatment procedures are also prearranged to the views of medical practitioners for dynamic and persuasive treatment of the patients.

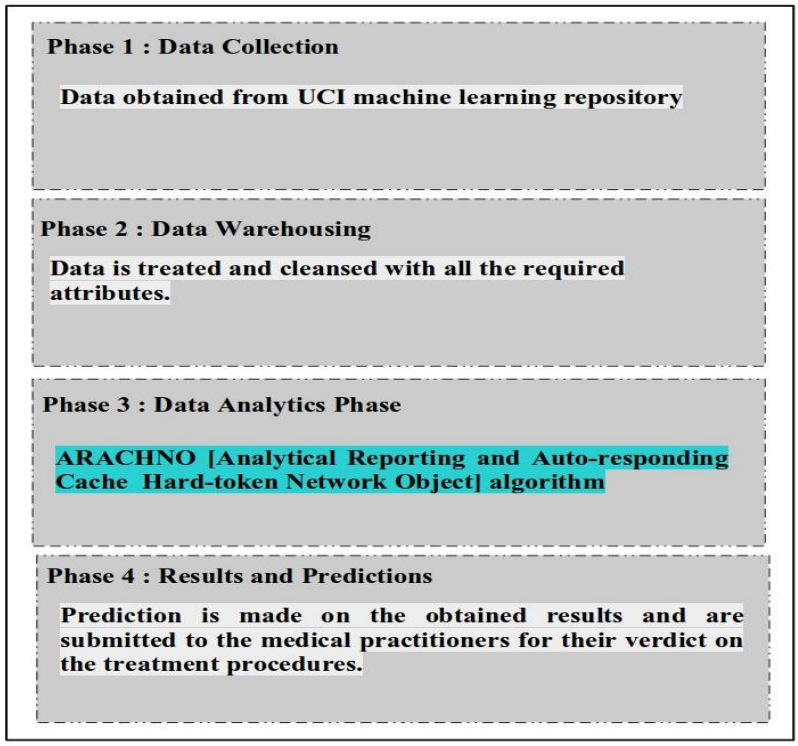

Figure 2: Proposed Predictive Methodology Architecture for executing ARACHNO algorithm

Algorithm Name: Analytical Reporting and Auto-responding Cache Hard-token Network Object [ARACHNO] algorithm

---

Input: Diabetic Patients data-intensive
Output: Predictive report

// Load the database with added medical reports [HbA_{1c} factor]

// Construct a data pattern by organizing the data-sets into structured order

1. Database = Order-transact (age,location,glucose level,code indicator)

// Construct a cache tree with the assembled structural database

2. Analytical-cache Tree= get Analytical-cache (Transaction Database)

// Assign a Hard-Token based on the server they are loaded

3. Hard-token= Order-transact(transaction);

// Construction of frequent pattern

4. Frequent-pattern = Order-transact(Analytical-cache Tree)

// Value assignment to the processed data

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

// Age factor, Glucose level factor and medication code as assigned in the data-sets are considered and the token value is also considered

5. **Hard-token number**= Order-transact(transaction_as_stored)
6. Order-transact(transaction)_calc = [age criteria + glucose level + medication code]
7. **Object_new_value** = {Hard-token number + Order-transact(transaction)_calc }

//All the Conditional criteria(cndtn) are defined

8. **Pre_def_cndtns**= [cndtn-1; cndtn-2; cndtn-3; cndtn-4; cndtn-5]

//All the reporting and medication conditions are defined

9. **report_object**= [rept-type1; rept-type-2; rept-type-3; rept-type-4; rept-type-5; rept-type-6; rept-type 7; rept-type 8; rept-type 9; rept-type 10]
10.**med_prac_cndtns**= [med_cndtn-1; med_cndtn-2; med_cndtn-3; med_cndtn-4; med_cndtn-5; med_cndtn-6; med_cndtn-7;]

//Prediction is made by comparing the condition value obtained with the predefined pattern set of prediction

11. ***Prediction score value*** [$Pred\_Score\_Val_{rep}$] is calculated
12. **if** (Object_new_value == Pre_def_cndtns) **do**
13. Compute med_prac_value based on med_prac_cndtns [1- 7]
14.**if** med_prac_value are higher than predefined threshold then
15. Hard-token number( ) ← Hard-token number (Hard-token number( ))
16. Increase Analytical-cache Tree numbering
17. Restart Algorithm

IV. EXPERIMENTAL RESULTS

The diabetic patients data-sets are obtained from the UCI machine learning repository. The condition of the diabetic patients can be classified into four main categories based on which the medication procedures can be prescribed. These conditions are Normal Glucose resilient stage, Insulin discretionary stage, Insulin required for regulation stage and Insulin required for subsistence stage. Glycated hemoglobin (HbA_{1c}) is a com-positional type of hemoglobin that is used to measure the plasma glucose concentration. However, since Red Blood Corpuscles do not all undergo lysis at the same time, Glycated hemoglobin (HbA_{1c}) is taken as a measure with limitation attribute and is taken to a time period of 3 months. The Analytical Reporting and Auto-responding Cache Hard-token Network Object algorithm produces a 'risk score' ranging from 0 to 5 for each of the uploaded data-sets. The output of this Analytical Reporting and Auto-responding Cache Hard-token Network Object algorithm is detailed to the extent where not only the diabetes type is identified but also the medication prescriptions along with treatment procedures are also prearranged to the views of medical practitioners for dynamic and persuasive treatment of the patients. The code field can be interpreted as given in table 2.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

Prediction score value [$Pred\_Score\_Val_{rep}$]	**Interpretation of the value**
0	Less possibility of Re-admittance
1	May be prescribed change of medication
2	Must be carefully monitored as the risk percentage is 30 to 45%
3	More possibility of Re-admittance as the risk percentage is 50 to 65%
4	More possibility of Re-admittance as the risk percentage is nearly 80%
5	Strictly Danger mode. Requires High precautionary measures of medication with close watch on all aspects

Table 2: Prediction Score Value and its interpretation

V. CONCLUSION AND FUTURE WORK

The proposed system for predicting the apt medication procedures for diabetic patients is implemented using Hadoop and Map-reduce paradigm. The method of using Glycated hemoglobin (HbA_{1c}) improves the output result. In future, this algorithm can be made to predict the onset of Diabetes mellitus at an early stage and can be made as an input for improving the medication procedures for the diabetic patients.

REFERENCES

[1] Frost, D.W., Vembu, S., Wang, J., Tu, K., Morris, Q. and Abrams, H.B., 2017. Using the Electronic Medical Record to Identify Patients at High Risk for Frequent ED Visits and High System Costs. *The American Journal of Medicine*.
[2] Liu, Yi, Yinghui Zhang, Jie Ling, and Zhusong Liu. "Secure and fine-grained access control on e-healthcare records in mobile cloud computing."*Future Generation Computer Systems* (2017).
[3] Turvey, Carolyn. "Personal Health Records, Patient Portals, and Mental Healthcare." In *Career Paths in Telemental Health*, pp. 115-121. Springer International Publishing, 2017.
[4] Nimmesgern E, Benediktsson I, Norstedt I. Personalized Medicine in Europe. Clinical and translational science. 2017 Mar 1;10(2):61-3.
[5] Wang, Y. and Hajli, N., 2017. Exploring the path to big data analytics success in healthcare. *Journal of Business Research*, *70*, pp.287-299.
[6] Wu, Po-Yen, Chih-Wen Cheng, Chanchala D. Kaddi, Janani Venugopalan, Ryan Hoffman, and May D. Wang. "–Omic and Electronic Health Record Big Data Analytics for Precision Medicine."*IEEE Transactions on Biomedical Engineering* 64, no. 2 (2017): 263-273.
[7] Nelson, Ramona, and Nancy Staggers. *Health informatics: An interprofessional approach*. Elsevier Health Sciences, 2017.
[8] Leyens L, Reumann M, Malats N, Brand A. Use of big data for drug development and for public and personal health and care. Genetic Epidemiology. 2017 Jan 1;41(1):51-60.
[9] World Health Organization. *Global report on diabetes*. World Health Organization, 2016.
[10] Mathers, Colin D., and Dejan Loncar. "Projections of global mortality and burden of disease from 2002 to 2030."*Plos med* 3, no. 11 (2006): e442.
[11] Samandari, N., Mirza, A.H., Nielsen, L.B., Kaur, S., Hougaard, P., Fredheim, S., Mortensen, H.B. and Pociot, F., 2017. Circulating microRNA levels predict residual beta cell function and glycaemic control in children with type 1 diabetes mellitus. *Diabetologia*, *60*(2), pp.354-363.
[12] Zhao, Danqing, et al. "Identification of candidate biomarkers for the prediction of gestational diabetes mellitus in the early stages of pregnancy using iTRAQ quantitative proteomics."*PROTEOMICS-Clinical Applications* (2017).
[13] Li, Y.H., Zhang, G.G. and Wang, N., 2017. Systematic Characterization and Prediction of Human Hypertension GenesNovelty and Significance. *Hypertension*, *69*(2), pp.349-355.
[14] Songthung, Phattharat, and Kunwadee Sripanidkulchai. "Improving type 2 diabetes mellitus risk prediction using classification." In *Computer Science and Software Engineering (JCSSE), 2016 13th International Joint Conference on*, pp. 1-6. IEEE, 2016.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

[15] Anand, T., Pal, R. and Dubey, S.K., 2016, March. Data mining in healthcare informatics: Techniques and applications. In *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on* (pp. 4023-4029). IEEE.
[16] Groves, P., Kayyali, B., Knott, D. and Kuiken, S.V., 2016. The big data revolution in healthcare: Accelerating value and innovation.
[17] DuBrava S, Mardekian J, Sadosky A, Bienen EJ, Parsons B, Hopps M, Markman J. Using Random Forest Models to Identify Correlates of a Diabetic Peripheral Neuropathy Diagnosis from Electronic Health Record Data. Pain Medicine. 2016 May 30:pnw096.
[18] Shetty SP, Joshi S. A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique. International Journal of Information Technology and Computer Science (IJITCS). 2016 Nov 8;8(11):26.
[19] Perveen, S., Shahbaz, M., Guergachi, A. and Keshavjee, K., 2016. Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*, *82*, pp.115-121.
[20] Crooks CJ, Card TR, West J. The use of a bayesian hierarchy to develop and validate a co-morbidity score to predict mortality for linked primary and secondary care data from the NHS in England. PLoS One. 2016 Oct 27;11(10):e0165507.
[21] White T., May 2012, "Hadoop: The Definitive Guide", Third Edition, O'Reilly, 978-1-449-31152-0
[22] Saecker M. and Markl V., "Big Data Analytics on Modern Hardware Architectures: A Technology Survey", 2013, Springer Lecture Notes in Business Information Processing, Volume 138, pp 125-149
[23] J. Dean and S. Ghemawat, "Mapreduce: a flexible data processing tool," Communications of the ACM, vol. 53, no. 1, pp. 72–77, 2010.
[24] R. C. Taylor, "An overview of the Hadoop/MapReduce/Hbase framework and its current applications in bioinformatics. ," BMC Bioinformatics, vol. 11, no. 12,2016.
[25] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.