Applications of Big Data Analytics for Diagnosing Diabetic Mellitus: Issues and Challenges

S.Shyni¹, R. Shantha Mary Joshitta², Dr. L. Arockiam³
¹M.Phil Scholar, ²Ph.D Scholar, ³Associate Professor,
¹²³Department of Computer Science, St.Joseph’s College, Tiruchirappalli, TamilNadu, India

Abstract: Today, the world has been excited about big data and the advancement in analytics. Big Data has an incredible influence from global market to the day-today life. Big Data is the term used to describe the huge volume of both structured and unstructured data. Big Data Analytics is the use of advanced analytical techniques against very large and diverse data sets that include different types and sizes. Nothing is more important than healthy and peaceful life. Big Data Analytics improves health care system through the reduction run time and the optimal cost. This paper presents an overview of big data analytics in health care and discusses the issues and challenges.

Keywords: Bigdata, Big data analytics, Healthcare.

I. INTRODUCTION

Healthcare environment is navigating today in a rapid speed. Though there are many challenges, there are few bright spots on the horizon by the development of technology. Reducing health care costs while improving patient care is an ongoing issue for the healthcare environment. In recent days, healthcare industry faces vast challenges to save the data generated and to process them to drive knowledge out of them. Big data analytics is the process of examining the large sets of data. Through Big data analytics, valuable insights can be gained, novel patterns can be detected, and powerful real world decisions can be made. Big Data Analytics permits data scientists and various other users to assess large volumes of transaction data.

II. BIG DATA

Big-data is nothing but a data available at autonomous and heterogeneous sources in extreme large amount which gets updated within a fraction of second [1].The characteristics of big data are:

a) Volume: It refers to the size of the data.
b) Velocity: It refers to the speed of data generation.
c) Variety: It refers to the different type of data being generated. The format of data may be structured, semi-structured or unstructured.
d) Value: It refers to the correctness of data.
e) Veracity: It refers to the trustworthiness of the data.
f) Variability: It refers to the constantly changing nature of Big Data.
This big data can be classified into three types. They are: (a) Structured data (b) Unstructured data and (c) Semi-Structured data.

a. **Structured data**: Data is in a definite format. Examples of structured data are spreadsheet, Relational databases etc.

b. **Unstructured data**: Data does not follow a particular layout. Examples of unstructured data include images, audio, document etc.

c. **Semi-Structured data**: Data Stored not in fixed field but the data includes metadata, schema or etc. Example: HTML etc.

1. **BIG DATA ANALYTICS**

Big data analytics is the process of collecting, organizing and analyzing large amount of data (called big data) to determine patterns and other valuable information. It can be classified into four types. They are:

a. **Prescriptive analytics** is important, but mostly not used.

b. **Predictive analytics** use big data to identify past patterns to predict the future.

c. **Diagnostic analytics** is used to discovery or to determine why something happened

d. **Descriptive analytics** or data mining are at the bottom of the big data value chain, but they can be valuable for uncovering patterns that offer insight.

III. **REVIEW OF LITERATURE**

Raghunath Nambiar et al. described the Big Data overview, the applications and the improvement in big data analytics [2]. The combination of real time services, people, clinical systems and historical population data made Big Data very helpful tool in improving the health care system. They pointed out the future scope of big data in healthcare.
Shantha Mary Joshitta et al. described a predictive model to estimate the diabetes using clinical big data in cloud [3]. The authors explained the need for analytical solutions in healthcare management and proposed a predictive model for streaming of data for analysis. The authors implemented the pattern generator and pre-processing system through pseudo code. The outcome of the analysis resulted in decrease of cost and diminished hospital readmissions.

Marco Viceconti et al. highlighted the importance of combining Big Data Analytics combined with Virtual Physiological Human (VPH) technologies to generate in silico medicine solutions [4]. For this purpose, the existing Big Data Technologies had been optimized to meet the specific need of this application. VPH model necessitated the development of mathematical model to predict events happening in the biological system. This challenge could be tackled through big data imaging and sensing technology. The specific needs included analytics of heterogeneous and sensible data ensuring security and performance constraints enhanced analytics to combine bio clinical observations. Through targeted funding all the requirements had been met using Big Data technologies.

Muni Kumar et al. illustrated the deficiency of suitable healthcare services and addressed the opportunity for healthcare services in rural India [5]. The e-health was mentioned not only to find the patient details but also medical reports. It was also for examining the patients for any hospitals (or) home (or) while on moving through mobiles. It gave better results at very low of cost.

Prajna Dora et al. illustrated the fraud activities in the field of health insurance [6]. They detected the doubtful records by implementing various analytic modules like decision tree, clustering etc. They pointed out many future directions and different analytic methods for detection of fraud.

Gomathi et al. highlighted the application of big data analytics and data mining in healthcare sector [7]. The challenges faced while using Big data were infrastructure security, data privacy, data management.

Peter Augustine outlined the benefits of big data analytics and Hadoop in the applications of healthcare [8]. The author expressed that the unstructured format of the data was the most important barrier for capturing the images when the surgery was going on. He used Hadoop Image Processing Interface (HIPI) model for processing images in the distributed computing environment. This concept delivered health services to everyone in moderate cost.

Rebecca Hermon et al. suggested that big data was very useful in healthcare [9]. The proposed healthcare information system reduced the cost of healthcare and improved the patient health outcomes.

SaumyaSalian et al. described the risk of re-admission for diabetic patients [10]. The authors analyzed the diabetic patients using decision tree. This decision tree was applied to know the patient’s diabetic condition. The proposed work was helpful to reduce the diabetic patients’ re-admission.

Ahmed E. Youssef pointed out security on health care systems with the help of Big Data Analytics in Mobile Cloud Computing Environments [11]. The author introduced a new framework for sharing of Electronic Health Records to Health Care Provider. They pointed out an important challenge, leaking the patient details through cloud. At the same time, he pointed out the loss of patients’ details and illegal usage of them by unauthorized persons. It helped the doctors to guide the patients at correct time. It provided high quality of healthcare in optimal cost.
SaravanaKumar N.M et al. illustrated the predictive methodology for Diabetic Data Analysis in Big Data [12]. In this paper, they described the Predictive analysis algorithm in Hadoop/MapReduce environment to predict the diabetes. They described the predictive analysis system in various phases like Data collection, data warehousing, predictive analysis and pre-processing. The analyzed reports in predictive level system helped to detect diseases at preliminary stages and provided the medical facility at minimum cost. It reduced and saved the next generation from diabetic mellitus.

WullianallurRaghupathi et al. pointed out the review of big data analytics in healthcare [13]. The architectural framework, platforms, tools and methodologies were mentioned in this paper.

Tae-Woongkim et al. gave an idea about the big data framework for u-healthcare system [14]. They explained the u-health care system. In u-health care system, vital signs depended on the two types of data. They were distinguishable data type and continuous linear structure data type. Variety of analytical techniques for big data was reputation analysis, social network analysis and etc. In this framework, the algorithm extracted the values of vital signs and Service Oriented Architecture for analysis and results. These methods solved lack of interoperability and systems integration problem.

Vijay M.Mane et al. explained the automated retinal image processing to detect diabetic retinopathy and presented few pre-processing methods to enhance the quality of the retinal images [15]. The method minimized the image dissimilarity and improves image quality. The proposed pre-processing technique normalized the original retinal image against a reference model. They explained the blood vessels extraction using K-means clustering and Fuzzy C-means clustering. The performance of preprocessing algorithm was evaluated using Peak Signal to Noise Ratio and Mean Square Error.

AmatulZehra et al. received several data mining applications, applied on the PIMA Indian Diabetes Data Set [16]. Diagnosis of diabetes had been done by employing various data mining techniques such as classification, clustering, association and pattern deduction. This study explored various issues and contributing factors that subsequently led to the development of diabetes and defined the ranges of normal and high blood glucose levels through the architecture. This work had proved the classification accuracies of preprocessed data over non preprocessed data.

George Krempl et al. described the challenges in data stream mining research [17]. They mentioned the privacy and confidentiality in data streams. The challenges were noisiness, outliers and adaptive instance selection. The problem of missing values referred to incompleteness of features. The challenges for active learning posed by evolving data streams were uncertainty regarding convergence, necessity of perpetual validation, temporal budget allocation and performance bounds. Algorithms to Decision Support System faced two challenges. They were making models simpler more reactive and more specialized and dealing with legacy systems.

SrideivanaiNagarajan et al. designed and implemented the clinical system to detect diabetics and the risks associated with it using clustering and classification techniques [18]. The data was gathered using data collection tools such as questionnaire, interviews and surveys. The experimental data was clustered using Simple K-means. The clustered data has been classified using classification algorithm such as Naive Bayes, Random Tree and etc. Classification used in this work was two-step process namely model construction and model usage. Eventually the efficiency of clinical system had been proved using clustering and classification data mining techniques.
Nithya et al. described the performance of clustering technique for the diabetes data set [19]. This clustering technique belonged to an unsupervised learning. Datasets were used for the comparison of the clustering algorithms. They helped for finding the best algorithm for the diabetes dataset. The experimental measures calculated the performance factors such as accuracy and execution time. The algorithms were examined by using the training set parameter based on its class attributes. It was stated that K-means algorithm gave better performance when comparing with the other two clustering algorithms.

J.N.Mamman et al. pointed out the diabetes classification using cascaded data mining techniques [20]. The techniques made use of pre-processed data and returned knowledge as output. The author used two different blocks of block diagram in cascaded classification techniques. K-means clustering algorithm used to classify the objects based on attributes. In this paper four types of accuracies were used namely, False Negative, True Positive, True Negative, and False Positive.

Sadhana et al. explained the analysis of data set with the help of Hive and R tool [21]. The Pima Indian Diabetic Data Set was used in this analysis. The data set was examined using Hive and R. R was one of the best languages used for statistical computing as well as for creating graphs. Hive was a data warehousing solution built on top of Hadoop the data set loaded in Hive. The raw data loaded to hive represented through the snapshot. They calculated the Gini index for each attribute. The attributes could be analyzed using the graphs generated.

IV. ISSUES AND CHALLENGES

- **Scalability**: Scalability is nothing but rapid increase in size [22]. As data volume increases, the value of different data records will decrease. Data volume increases at a quicker speed than computing resources and CPU speed.
- **Data Complexity**: It is one of the current difficulties of big data [24]. It is also necessary to connect and correlate relationships, hierarchies and multiple data linkages or data can quickly twisting out of control.
- **Speed**: For Big Data, speed is an important issue[24]. This speed is also relevant to scalability. The speed of data generation depends on two factors: data access time and efficiency.
- **Accuracy**: It is not a serious issue. With the promise to big data, the data sources are of many origins, not at all verifiable [22]. Therefore the accuracy of the data becomes a serious issue.
- **Cost**: Big Data has opened up a world of possible business improvements; there is a great deal of experimentation and discovery taking place to determine the patterns that matter and the insights that turn to value [24]. To make sure a positive on a Big Data project, it is crucial to reduce the cost of the solutions used to find the value.
- **Storage**: Large and extensive big datasets must be managed with reliability, availability and easy accessibility [23]. Storage capacity can be competitive if it increases in volume and storage. Hence research on data storage is necessary.
- **Security**: Security concerns data protection. It is a major obstacle preventing companies from taking full advantage of their data [23]. There should be a balance between data privacy and national security.
- **Processing**: Processing large amount of data takes a lot of time. Research is needed to decrease processing time.
## V. A STUDY ON ISSUES AND CHALLENGES

<table>
<thead>
<tr>
<th>Author</th>
<th>Title of the paper</th>
<th>Issue Identified</th>
<th>Tool/ Algorithm Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muni Kumar [5] et al.</td>
<td>Role of Big Data Analytics in Rural Health Care. A step towards SvasthBharath</td>
<td>Provided e-health to rural areas</td>
<td>HDFS</td>
</tr>
<tr>
<td>M.Gomathi [7] et al.</td>
<td>Applications of Big Data Analytics and Data Mining in HealthCare Sector</td>
<td>Maintaining the security of data, Storage</td>
<td>Hadoop</td>
</tr>
<tr>
<td>D.Peter Augustine[8]</td>
<td>Leveraging Big Data Analytics and Hadoop in Developing India’s Healthcare Services</td>
<td>Different formats of Data</td>
<td>Hadoop</td>
</tr>
<tr>
<td>Rebecca Hermon [9] et al.</td>
<td>Big data in healthcare: What is it used for?</td>
<td>Cost</td>
<td>-</td>
</tr>
<tr>
<td>Authors</td>
<td>Title</td>
<td>Cost</td>
<td>Technologies</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>-------------------------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Dr. Saravana Kumar et al.</td>
<td>Predictive Methodology for Diabetic Data Analysis in Big Data</td>
<td>Cost</td>
<td>Hadoop, Map Reduce</td>
</tr>
<tr>
<td>Tae-Woong Kim et al.</td>
<td>Big Data Framework for u-healthcare System</td>
<td>Data Interoperability and System Integration</td>
<td>Hadoop</td>
</tr>
<tr>
<td>Vijay M. Mane et al.</td>
<td>Preprocessing in Early Stage Detection of Diabetic Retinopathy Using Fundus Images</td>
<td>Segmentation of retinal blood vessels</td>
<td>K-Means clustering and Fuzzy C means clustering</td>
</tr>
<tr>
<td>Amatul Zehra et al.</td>
<td>A comparative study on the pre-processing and mining of Pima Indian Diabetes Dataset</td>
<td>Model to predict and diagnose diabetes</td>
<td>WEKA</td>
</tr>
<tr>
<td>Sridevi Nagarajan et al.</td>
<td>Design and Implementation of Expert Clinical System for Diagnosing Diabetes using Data Mining Techniques</td>
<td>Predicting the type of diabetes and risk levels of diabetes</td>
<td>Classification and Clustering</td>
</tr>
<tr>
<td>R. Nithya et al.</td>
<td>Analysis of clustering technique for the diabetes dataset using the training set parameter</td>
<td>Accuracy and execution time</td>
<td>Clustering</td>
</tr>
<tr>
<td>J. N. Mamman et al.</td>
<td>Diabetes classification using Cascaded Data Mining Techniques</td>
<td>Accuracy</td>
<td>K-means-ANN model</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

This study reviewed the already existing literature on healthcare. In this paper, various issues and challenges in big data analytics in healthcare were highlighted. The techniques discussed here provide motivation for applying big data analytics in the healthcare domain. In future, big data analytics will spread in different areas of healthcare industry.

REFERENCES