© 2018 IJSRST | Volume 4 | Issue 2 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

A Review on Big Data Application in Health Care

Sridhar Gujjeti[1], Dr. Suresh Pabboju[2]
[1]Assistant Professor, Department of Computer Science and Engineering, Kakatiya Institute of Technology and Science, Warangal, Telangana, India
[2]Professor, Department of Information Technology, Chaitanya Bharathi Institute of Technology(CBIT), Hyderabad, Telangana, India

ABSTRACT

Big data technologies are progressively utilized for biomedical and health-care informatics research. A lot of biological and clinical data have been created and gathered at a phenomenal speed and scale. For instance, the new age of sequencing technologies empowers the handling of billions of DNA sequence data every day, and the application of electronic health records (EHRs) is archiving a lot of patient data. The cost of getting and breaking down biomedical data is required to diminish drastically with the assistance of innovation redesigns, for example, the rise of new sequencing machines, the advancement of novel equipment and programming for parallel computing, and the broad extension of EHRs. Big data applications introduce new chances to find new information and make novel strategies to enhance the nature of health care. The application of big data in health care is a quickly developing field, with numerous new disclosures and philosophies distributed over the most recent five years. In this paper, we review and talk about big data application in four noteworthy biomedical sub disciplines: (1) bioinformatics, (2) clinical informatics, (3) imaging informatics, and (4) general health informatics. In particular, in bioinformatics, high-throughput tests encourage the research of new expansive affiliation investigations of diseases, and with clinical informatics, the clinical field benefits from the immense measure of gathered patient data for settling on smart choices. Imaging informatics is presently more quickly incorporated with cloud stages to share medical image data and work processes, and general health informatics use big data methods for foreseeing and observing infectious disease flare-ups, for example, Ebola. In this paper, we review the current advance and achievements of big data applications in these health-care domains and condense the difficulties, holes, and chances to enhance and progress big data applications in health care.

Keywords: Big Data, Literature Review, Health Care, Data-Driven Application

I. INTRODUCTION

In the biomedical informatics domain, big data is another worldview and a biological system that changes case-based investigations to huge scale, data-driven research. It is generally acknowledged that the qualities of big data are characterized by three noteworthy highlights, ordinarily known as the 3Vs: volume, assortment, and speed. To start with and most essentially, the volume of data is developing exponentially in the biomedical informatics fields. For instance, the ProteomicsDB8 covers 92% (18,097 of 19,629) of known human qualities that are explained in the Swiss-Prot database. ProteomicsDB has a data volume of 5.17 TB. In the clinical domain, the advancement of the HITECH Act9 has about tripled the reception rate of electronic health records (EHRs) in doctor's facilities to 44% from 2009 to 2012. Data from a great many patients have just been gathered and put away in an electronic arrangement, and these collected data could possibly improve health-care administrations and increment research opportunities.10,11 what's more, medical imaging (eg, MRI, CT filters) produces tremendous measures of data with considerably more perplexing highlights and more

extensive measurements. One such case is the Visible Human Project, which has documented 39 GB of female datasets.12 These efficient devices for finding new examples among populace bunches utilizing web-based social networking data.

II. BIG DATA TECHNOLOGIES

Biomedical scientists are confronting new difficulties of putting away, overseeing, and investigating monstrous measures of datasets. The attributes of big data require capable and novel technologies to extricate helpful data and empower more wide based health-care arrangements. In the greater part of the cases revealed, we found various technologies that were utilized together, for example, artificial intelligence (AI), alongside Hadoop®, and data mining devices. Parallel computing is one of the central foundations for overseeing big data errands. It is equipped for executing calculation undertakings at the same time on a group of machines or supercomputers. As of late, novel parallel computing models, for example, MapReduce by Google, have been proposed for another big data framework. All the more as of late, an open-source MapReduce package called Hadoop was released by Apache for distributed data management. The Hadoop Distributed File System (HDFS) supports concurrent data access to clustered machines. As such, cloud computing is a novel model for sharing configurable computational resources over the network and can serve as an infrastructure, platform, and/or software for providing an integrated solution. Many new big data applications are based on cloud technologies.

III. RESEARCH METHODS

We sought four bibliographic databases to discover related research articles: (1) PubMed, (2) ScienceDirect, (3) Springer, and (4) Scopus. In looking through these databases, we utilized the fundamental catchphrases "big data," "health care," and "biomedical." Then, we chose papers in light of the accompanying consideration criteria:

1. The paper was composed in English and distributed inside the previous five years (2000–2015).
2. The paper talked about the outline and utilization of a big data application in the biomedical and health-care domains.
3. The paper detailed another pipeline or strategy for handling big data and talked about the execution of the technique.
4. The paper assessed the execution of new or existing big data applications.

The accompanying avoidance criteria were utilized to sift through immaterial papers:

1. The paper did not examine a particular big data applications (eg, general remarks about big data).
2. The paper was an instructional exercise or a course material.
3. The paper was not in the four concentration territories: bioinformatics, clinical informatics, general health informatics, and imaging informatics.

Two hunts were performed. In the principal look, the primary author (JL) and the second author (MW) of the present investigation started the inquiry procedure in view of the principle watchwords. All conceivably related papers were gathered by reviewing the title and unique. This underlying inquiry brought about 755 papers from 2000 to 2015. In the second pursuit, the second author (MW) and the third author (DG) screened the papers in light of the previously mentioned consideration and prohibition criteria and hence chose 94 candidate papers. At long last, each author of the present examination assessed the last determination by perusing the substance of the papers, and agreement was come to review 68 papers for this investigation.

IV. BIG DATA APPLICATIONS

Bioinformatics applications. Bioinformatics research dissects biological framework variations at the molecular level. With current patterns in customized solution, there is an expanding need to create, store, and break down these huge datasets in a reasonable time period. Next-generation sequencing innovation empowers genomic data obtaining in a brief timeframe. The part of big data methods in bioinformatics applications is to give data vaults, computing foundation, and productive data manipulation apparatuses for investigators to gather and break down biological information. Taylor examines that Hadoop and MapReduce are presently utilized broadly inside the biomedical field.

Big data technologies are divided into four categories : (1) data storage and retrieval, (2) error identification, (3) data analysis, and (4) platform integration deployment. These categories are correlated and may cover; for example, most data input applications may bolster straightforward data analysis, or the other way around. Be that as it may, our classification in the present examination is construct just with respect to the fundamental elements of every innovation.

Data storage and retrieval. These days, a sequencing machine can deliver a great many short DNA sequencing data amid one run. The sequencing data should be mapped to particular reference genomes with a specific end goal to be utilized for extra analysis, for example, genotype and articulation variation analysis. Downpour is a parallel computing model that facilitates the genome mapping process. Torrent parallelizes the short-read mapping procedure to enhance the versatility of perusing extensive sequencing data. The CloudBurst demonstrate was evaluated utilizing a 25-center bunch, and the outcomes indicate that the speed to process seven million short-peruses was just about 24 times quicker than a solitary center machine. The CloudBurst group have grown new instruments in view of CloudBurst to help biomedical research, for example, Contrail for collecting vast genomes and Crossbow for distinguishing single nucleotide polymorphisms (SNPs) from sequencing data.

Clinical informatics applications. Clinical informatics centers around the application of information innovation in the health-care domain. It incorporates movement based research, analysis of relationship between patient principle determination (MD) and hidden reason for death (UCD), and storage of data from EHRs and different sources (eg, electrophysiological [such as EEG] data). In this area, we ordered big data technologies/devices into four categories: (1) data storage and retrieval, (2) interactive data retrieval for data sharing, (3) data security, and (4) data analysis. Contrasted and bioinformatics, clinical informatics does not offer numerous instruments for error identification but rather gives careful consideration to data-sharing and data security issues. Its data analysis strategy is altogether different from bioinformatics, as clinical informatics works with both organized and unstructured data, creates particular ontologies, and utilizations natural dialect preparing widely.

Data storage and retrieval. It is basic to talk about the manners by which big data systems (eg, Hadoop, NoSQL database) are utilized for putting away EHRs. The proficient storage of data is particularly imperative when working with clinical ongoing stream data. Dutta et al evaluated the capability of utilizing Hadoop and HBase as data distribution centers for putting away EEG data and talked about their high-performance qualities. Jin et al. broke down the capability of utilizing Hadoop HDFS and HBase for appropriated EHRs. Besides, Sahoo et al. and Jayapandian et al. proposed an appropriated structure for putting away and questioning a lot of EEG data. Their framework, Cloudwave, utilizes Hadoop-based data preparing modules to store clinical data, and by

utilizing the handling energy of Hadoop, they built up an electronic interface for ongoing data visualization and retrieval. The Cloudwave group evaluated a dataset of 77-GB EEG flag data and contrasted Cloudwave and a stand-alone framework; the outcomes demonstrate that Cloudwave prepared five EEG thinks about in 1 minute, while the stand-alone framework took over 20 minutes. Contrasted and a customary relational database that handles organized data well, the novel NoSQL is a great prospect for putting away unstructured data. Mazurek proposed a framework that joins both relational and multidimensional technologies with NoSQL storehouses to empower data mining systems and give adaptability and speed in data handling. Nguyen et al. exhibited a model framework for putting away clinical flag data, where the time arrangement data of clinical sensors are put away inside HBase in a way that the line key fills in as the time stamp of a solitary esteem, and the segment stores patient physiological esteems that relate with the line key time stamp. To enhance the availability and read-capacity of the HBase data mapping, the metadata are put away in MongoDB, which is a report based NoSQL database. Google Web Toolkit is incorporated into the framework to picture the clinical flag data.

Public health information. As depicted by Short-liffe and Cimino, public health has three center capacities: (1) assessment, (2) policy development, and (3) assurance. Among these, evaluation is the essential and fundamental capacity. Evaluation essentially includes gathering and breaking down data to track and screen public health status, along these lines giving proof to basic leadership and policy development. Assurance is utilized to validate whether the ser-indecencies offered by health foundations have accomplished their underlying target objectives for expanding public health results; all things considered, numerous huge public health organizations, for example, the Centers for Disease Control and Prevention and the Administration of Community Living, have gathered and broke down a lot of population health data. In this segment, no new methodologies are presented. Rather, we show an integrated perspective of big data and health from a population point of view rather than a solitary medical/clinical movement viewpoint. This segment centers around four territories:
(1) infectious disease surveillance, (2) population health management, (3) mental health management, and (4) chronic disease management.

Infectious disease surveillance. Roughage talked about the open doors for utilizing big data for worldwide infectious disease surveillance. They built up a framework that gives continuous risk checking on delineate, out that machine learning and group sourcing have opened new conceivable outcomes for building up a persistently updated atlas for disease observing. Feed et al trusted that online web-based social networking joined with epidemiological information is a significant new data hotspot for facilitating public health surveillance. The utilization of web-based social networking for disease observing was demonstrated by Young et al., in which they gathered 553,186,016 tweets and separated more than 9,800 with HIV risk-related catchphrases (eg, sexual practices and medication utilize) and geographic annotations. They demonstrated that there is a huge positive correlation (P , 0.01) between HIV-related tweets and HIV cases in light of predominance analysis, illustrating the significance of online networking (eg, Twitter, Facebook) and its potential effect on checking worldwide disease event.

Population health management. To think about the appropriation and effect of sociodemographic and medico-administrative variables, Lamarche-Vadel et al. broke down the free association of patient MD and UCD. The MD was distinguished by ICD10 code, while the UCD was separated from a death registry. In the event that MD and UCD were diverse occasions, at that point those occasions were observed to be

autonomous. Utilizing health protection data, information from 421,460 perished patients was removed from 2008 to 2009. The outcomes demonstrate that 8.5% of in doctor's facility deaths and 19.5% of out-of-healing center deaths were free occasions and that autonomous death was more typical in elderly patients. The outcomes demonstrate that expansive scale data analysis can be utilized to adequately dissect the association of medical occasions.

Mental health management. Nambisan et al. discovered that messages posted via web-based networking media could be utilized to screen for and conceivably distinguish sadness. Their analysis depends on past research of the association between depressive issue and monotonous musings/ruminating conduct. Big data examination devices assume an imperative part in their work by mining shrouded behavioral and passionate patterns in messages, or "tweets," posted on Twitter. Inside these tweets, we might have the capacity to distinguish a disease-related feeling pattern, which is a formerly concealed manifestation. The authors foresee that future research could dive further into the conversations of the discouraged clients to understand more about their concealed feelings and notions. What's more, Dabek and Caban introduced a neural system demonstrate that can anticipate the probability of creating mental conditions, for example, nervousness, behavioral scatters, melancholy, and post-traumatic pressure issue. They additionally investigated the adequacy of their model against a dataset of 89,840 patients, and the outcomes demonstrate that they can accomplish a general exactness of 82.35% for all conditions.

Chronic disease management. Tu et al. presented the Cardiovascular Health in Ambulatory Care Research Team (CANHEART), a one of a kind, population-based observational research initiative went for estimating and enhancing cardio-vascular health and the nature of ambulatory cardiovascular care gave in Ontario, Canada. The research concentrated on recognizing chances to enhance the essential and optional avoidance of cardiovascular occasions in Ontario's assorted multiethnic population. The examination included data from

9.8 million Ontario grown-ups matured $20 years. Data were gathered by connecting numerous databases, for example, electronic reviews, health administration, clinical, laboratory, medicate, and electronic medical record databases utilizing encoded individual identifiers. Follow-up clinical occasions were gathered

V. CONCLUSION

We are right now in the period of "big data," in which big data innovation is in effect quickly connected to biomedical and health care fields. In this review, we demonstrated different cases in which big data innovation has assumed a vital part in cutting edge health-care upset, as it has totally changed individuals' perspective of health-care action. The initial three areas of this review uncovered that big data applications facilitate three imperative clinical activities, while the last segment (particularly the chronic disease management segment) draws an integrated picture of how separate clinical activities are finished in a pipeline to oversee singular patients from numerous viewpoints. We outlined late advance in the most pertinent zones in each field, including big data storage and retrieval, error identification, data security, data sharing and data analysis. Besides, in this review, we discovered that bioinformatics is the essential field in which big data examination are as of now being connected, generally because of the monstrous volume and multifaceted nature of bioinformatics data. Big data application in bio-informatics is relatively mature, with sophisticated platforms and apparatuses as of now being used to help investigate biological data, for example, quality sequencing mapping devices. Be that as it may, in other biomedical research fields, for example, clinical informatics, medical imaging informatics, and public

health informatics, there is huge, undiscovered potential for big data applications.

This literature review likewise demonstrated that: (1) integrating distinctive wellsprings of information empowers clinicians to delineate another perspective of patient care forms that think about a patient's all encompassing health status, from genome to conduct; (2) the benefit capacity of novel portable health technologies facilitates ongoing data gathering with more exactness; (3) the implementation of dispersed platforms empowers data filing and analysis, which will additionally be produced for choice help; and (4) the incorporation of land and environmental information may additionally build the capacity to translate gathered data and concentrate new learning.

While big data holds critical guarantee for enhancing health care, there are a few basic difficulties confronting all the four fields in utilizing big data innovation; the most huge issue is the integration of different databases. For instance, the VHA's database, VISTA, isn't a solitary framework; it is an arrangement of 128 interlinked frameworks. This turns out to be much more complicated when databases contain distinctive data composes (eg, integrating an imaging database or a laboratory test comes about database into existing frameworks), subsequently constraining a framework's capacity to make questions against all databases to gain every single patient datum. The absence of standardization for laboratory protocols and qualities additionally creates challenges for data integration. For instance, image data can experience the ill effects of innovative batch impacts when they originate from various laboratories under various proto-cols. Efforts are made to standardize data when there is a batch impact; this might be less demanding for image data, yet it is inherently more hard to standardize laboratory test data. Security and protection concerns additionally stay as obstacles to big data integration and utilization in all the four fields, and in this way, secure platforms with better communication standards and protocols are greatly required.

In its latest industry analysis report, McKinsey and Company anticipated that big data examination for the medical field will possibly spare more than $300 billion every year in US health-care costs. Future development of big data applications in the biomedical fields holds foreseeable guarantee since it is subject to the headway of new data standards, important research and innovation, cooperation in research foundations and organizations, and solid government motivations.

VI. REFERENCES

[1]. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458(7239):719-24.

[2]. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008; 26(10):1135-45.

[3]. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11(1):31-46.

[4]. Zhang J, Chiodini R, Badr A, et al. The impact of next-generation sequencing on genomics. J Genet Genomics. 2011;38(3):95-109.

[5]. Lynch C. Big data: how do your data grow? Nature. 2008;455(7209):28-9.

[6]. Wilhelm M, Schlegl J, Hahne H, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509(7502):582-7.

[7]. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med. 2010;363(6):501-4.

[8]. Shoban Babu Sriramoju, "Analysis and Comparison of Anonymous Techniques for Privacy Preserving in Big Data" in "International Journal of Advanced Research in Computer and Communication Engineering", Vol 6, Issue 12, December 2017, DOI 10.17148/IJARCCE.2017.61212 ISSN(online) : 2278-1021, ISSN(print) : 2319-5940

[9]. Shoban Babu Sriramoju, " Review on Big Data and Mining Algorithm" in "International Journal for Research in Applied Science and Engineering Technology", Volume-5, Issue-XI,

November 2017, 1238-1243 ISSN : 2321-9653], www.ijraset.com

[10]. Ackerman MJ. The Visible Human Project: a resource for education. Acad Med. 1999;74(6):667-70.

[11]. Feldman R, Sanger J., Advanced Approaches in Analyzing Unstructured Data. Cambridge: Cambridge University Press; 2007.

[12]. Xu H, Rosenbloom ST, Denny JC, Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc. 2011;18(2):181-6.

[13]. Weng C, Wu X, Luo Z, An approach for eligibility criteria extraction and representation. J Am Med Inform Assoc. 2011;18:i116-24.

[14]. Hanna M, McKenna A, Banks E. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.

[15]. Chou W-YS, Hunt YM, Beckjord EB, et al. Social media use in the United States: implications for health communication. J Med Internet Res. 2009;11(4):e48.

[16]. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis. 2009;49(10):1557-64.

[17]. Margolis R, Derr L, Dunn M, et al. The National Institutes of Health's Big Data to Knowledge (BD2 K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc. 2014;21(6):957-8.

[18]. White T. Hadoop: The Definitive Guide. Sebastopol, CA: O'Reilly Media, Inc.; 2012.

[19]. Shoban Babu Sriramoju, "Mining Big Sources Using Efficient Data Mining Algorithms" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol 2, Issue 1,January 2014 ISSN(online) : 2320-9801, ISSN(print) : 2320-9798

[20]. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM. 2008;51(1):107-13.

[21]. Armbrust M, Fox A, Griffith R, et al. A view of cloud computing. Commun ACM. 2010;53(4):50-8.

[22]. Schuster SC. Next-generation sequencing transforms today's biology. Nature. 2007;200(8):16-8.

[23]. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics. 2008;92(5):255-64.

[24]. Taylor R. An overview of the Hadoop/MapReduce/HBase framework and its cur- rent applications in bioinformatics. BMC Bioinformatics. 2010;11(suppl 12):S1.

[25]. Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce.Bioinformatics. 2009;25(11):1363-9.

[26]. Schatz M, Sommer D, Kelley D, et al. Contrail: assembly of large genomes using cloud computing. In: CSHL Biology of Genomes Conference, Cold Spring Harbor, New York; CSHL. 2010.

[27]. Gurtowski J, Schatz MC, Langmead B. Genotyping in the cloud with crossbow. Curr Protoc Bioinformatics. 2012;Chapter 15:Unit15.3.

[28]. Pandey RV, Schlötterer C. DistMap: a toolkit for distributed short read mapping on a Hadoop cluster. PLoS One. 2013;8(8):e72614.

[29]. George L. HBase: The Definitive Guide. Sebastopol, CA: O'Reilly Media, Inc.; 2011.

[30]. Huang W, Li L, Myers JR, et al. ART: a next-generation sequencing read simu- lator. Bioinformatics. 2012;28(4):593-4.

[31]. Shoban Babu Sriramoju, "OPPORTUNITIES AND SECURITY IMPLICATIONS OF BIG DATA MINING" in "International Journal of Research in Science and Engineering", Vol 3, Issue 6, Nov-Dec 2017 ISSN : 2394-8299.

[32]. Siripuri Kiran, 'Decision Tree Analysis Tool with the Design Approach of Probability Density Function towards Uncertain Data Classification', International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN : 2395-6011, Online ISSN : 2395-602X,Volume 4 Issue 2, pp.829-831, January-February 2018. URL : http://ijsrst.com/IJSRST1841198

[33]. Shoban Babu Sriramoju, Madan Kumar Chandran, "UP-Growth Algorithms for Knowledge Discovery from Transactional Databases" in "International Journal of Advanced Research in Computer Science and Software Engineering", Vol 4, Issue 2, February 2014 ISSN : 2277 128X

[34]. Amitha Supriya. "Implementation of Image Processing System using Big Data in the Cloud Environment." International Journal for Scientific Research and Development 5.10 (2017): 211-217.

[35]. SA Supriya. "A Survey Model of Big Data by Focusing on the Atmospheric Data Analysis." International Journal for Scientific Research and Development 5.10 (2017): 463-466.

[36]. Ajmera Rajesh, Siripuri Kiran, " Anomaly Detection Using Data Mining Techniques in Social Networking" in "International Journal for Research in Applied Science and Engineering Technology", Volume-6, Issue-II, February 2018, 1268-1272 ISSN : 2321-9653], www.ijraset.com

[37]. Dr. Shoban Babu Sriramoju, "A Review on Processing Big Data" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol-2, Issue-1, January 2014 ISSN(online) : 2320-9801, ISSN(print) : 2320-9798