Volume 6, Issue 10, October 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Big Data Analytics Research Opportunities and Challenges- A Review

G. Sabarmathi
Department of MCA, JNC, BU, Karnataka, India

Dr. R. Chinnaiyan
Department of MCA, NHCE, VTU, Karnataka, India

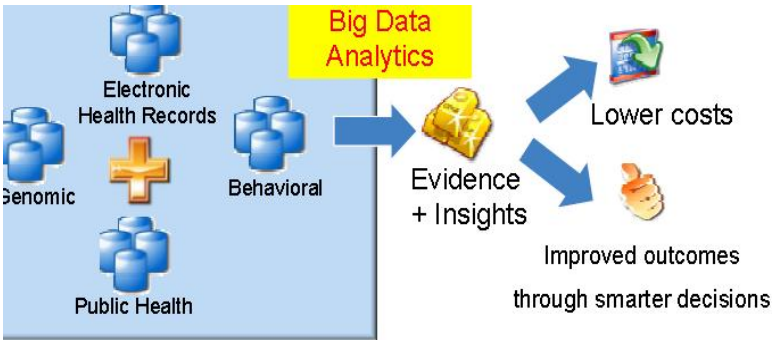
Dr. V. Ilango
Department of MCA, NHCE, VTU, Karnataka, India

***Abstract*— Even though big data technologies seem to be overhyped and promised to have great potential in the domain of medicine, if the development takes place in the integrated environment in combination with other modeling strategies, it will definitely going to ensure a unvarying enhancement of in-silico medicine and lead to favorable clinical adoption. This proposed research is planned to analyze the major issues in order to have an effective integration of big data analytics and efficient modeling in healthcare.**

***Keywords*— Big Data, Smart and Connected Communities (SCC), CDSS, EHR data**

I. INTRODUCTION

A collection of large and complex data sets which are difficult to processusing common database management tools or traditional data processing applications. Big data is not just about size. Finds insights from complex, noisy,heterogeneous, longitudinal, and voluminous data. It aims to answer questions that were previously unanswered. Big Data constantly facing significant challenges like outsized, heterogeneity, noisy labels, non-stationary distribution. capturing, storing,searching, sharing & analyzing. The four dimensions (V's) of Big Data It is important to recognize the full potential of Big Data by addressing these technical challenges with new ways of thinking and transformative solutions. If these challenges are resolved on time, there will be a plenteous opportunities to provide major advancement in science, medicine and business. While there is clearly an important research space examining the fundamental methods and technologies for big data analytics, it is vital to acknowledge that it is also necessary to fund domain-targeted research that allows specialized solutions to be developed for specific applications. Healthcare, in general, deserves to be a natural candidate for this kind of evaluation.



Above diagrammatic representation explains the advantage of the massive amounts of data which provide right intervention to the right patient at the right time. Personalized care to the patient that potentially benefit all the components of a healthcare system i.e., provider, payer, patient, and management.

II. LITERATURE REVIEW

M. Viceconti *et al*., [1] described five major problems in healthcare data management systems. These are as follows; 1. Working with sensitive Data. 2. Analytics of complex and heterogeneous data spaces, including nontextual information. 3. Distributed data management under security and performance constraints. 4. Specialized analytics to integrate bioinformatics and systems biology information with clinical observations at tissue, organ and organisms scales 5. Specialized analytics to define the "physiological envelope" during the daily life of each patient. J. Andreu-Perez *et al*., [2] provided an overview of recent developments in big data in the context of biomedical and health informatics. Yunchuan *et al*., [3] promoted the concept of ``smart and connected communities (SCC)", which is evolving from the concept of smart cities. SCC are envisioned to address synergistically the needs of remembering the past (preservation and revitalization), the needs of living in the present (livability), and the needs of planning for the future (sustainability). X. W. Chen and X. Lin [4] has given a brief overview of deep learning, and highlighted current research efforts and the challenges to big data, as well as the future trends. A. Fahad *et al*., [5] performed a survey on a comprehensive study of

the clustering algorithms proposed in the literature. In order to reveal future directions for developing new algorithms and to guide the selection of algorithms for big data, they proposed a categorizing framework to classify a number of clustering algorithms. The categorizing framework is developed from a theoretical viewpoint that would automatically recommend the most suitable algorithm(s) to network experts while hiding all technical details irrelevant to an application.

L. Xu *et al.*, [6] reviewed the privacy issues related to data mining by using a user-role based methodology. They differentiated four different user roles that are commonly involved in data mining applications, i.e. data provider, data collector, data miner and decision maker. A. Belle *et al*., [7] reviewed that the Big Data focused on three areas of interest: medical image analysis, physiological signal processing, and genomic data processing. V. Sujatha *et al*., [8] analyzed that the data sets from statistical models or complex pattern recognition models may be fused into predictive models that combines data set of patients' treatment information and prognostic outcome results. S. Vennila and J. Priyadarshini., [9] promoted that the security in Big data is a challenging research issue. If Integration of MapReduce, a machine for privacy preserving, is designed for the analyzing of data would provide better privacy.

Kovalchuk *et al*., [10] represented an early stage of the work aimed to the development of a general-purpose concept of the P4 CDSS rising from a treatment-level scope to a hospital-level scope. J. Cunha, C. Silvaa and M. Antunes [11] proposed a generic functional architecture with Apache Hadoop framework and Mahout for handling, storing and analyzing big data that can be used in different scenarios. Z. Liu *et al*., [12] presented an agent-based model of emergency department that was implemented in Netlogo simulation environment. Case studies have been carried out for proving two of the possible uses of the simulator, one to meet the increasing patient arrival overcrowding problem, and the second a quantitative analysis of the influence of ambulance response time (for departure) over the ED behavior.

M. Srivathsan and Y. Arjun [13] proposed that Prognotive Computing recognize patterns and formulates its own structure to provide a solution or gives a predicted alert so as to find a solution by ourselves .The System provides a handle of Health care and life span of numerous life forms. A. Abbas *et al*., [14] stated that they propose a cloud based framework that effectively manages the health related Big-data and benefits from the ubiquity of the Internet and social media. The framework facilitates the mobile and desktop users by offering: (a) disease risk assessment service and (b) consultation service with the health experts on Twitter. F. Zhang *et al*., [15] proposed a task-level adaptive MapReduce framework. This framework extends the generic MapReduce architecture by designing each Map and Reduce task as a consistent running loop daemon. The beauty of this new framework is the scaling capability being designed at the Map and Task level, rather than being scaled from the compute-node level. Y. Wang, L. Kung and T. A. Byrd [16] examined that health care industry has not fully grasped the potential benefits to be gained from big data analytics. K. Kambatla *et al*., [17] provided an overview of the state-of-the-art and focus on emerging trends to highlight the hardware, software, and application landscape of big-data analytics. J. Wang, M. Qiu and B. Guo [18] developed a telehealth system that covers both clinical and nonclinical uses, which not only provides store-and-forward data services to be offline studied by relevant specialists, but also monitors the real-time physiological data through ubiquitous sensors to support remote telemedicine. S. M. DeJong [19] proposed that technology is likely to become increasingly important in healthcare. Any professionalism concerns must be weighed against the potential benefits of technology to patients. P. Nadkarni [20] explained that the Institute of Medicine's idea of a learning health system, in which the boundaries between research and clinical practice are blurred.

The historical roots of this idea are identified by exploring initiatives in the business world such as knowledge management, business process reengineering, and enterprise resource planning. M. Legg [21] stated that the standardization required to achieve interoperability for pathology test requesting and reporting. Interoperability is the ability of two parties, either human or machine, to exchange data or information in a manner that preserves shared meaning. A. T. Janke *et al*., [22] explained that clinical research often focuses on resource-intensive causal inference, whereas the potential of predictive analytics with constantly increasing big data sources remains largely unexplored. Basic prediction, divorced from causal inference, is much easier with big data. L.A. Winters-Miner *et al*., [23] predicted the development of a healthcare-centered democracy and seen an explosion in the volume and velocity of patient-generated data. This development has become a driving force in the connection of digital health records to each other and to diagnosis and treatment practitioners.

III. RESEARCH OPPORTUNITIES AND CHALLENGES

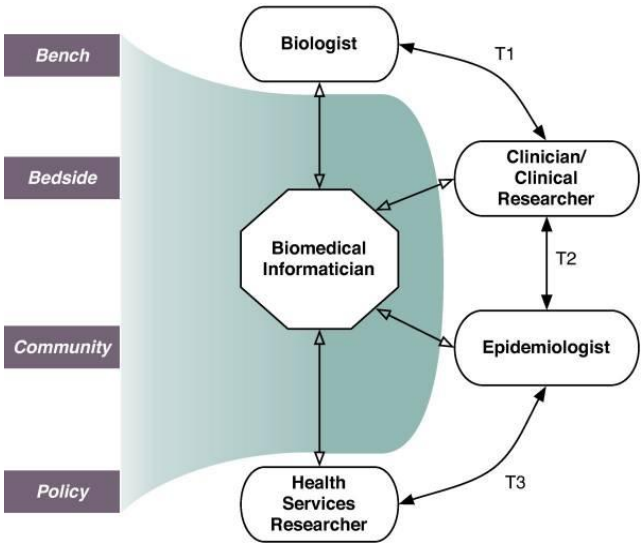
To analyze the major issues in order to have an effective integration of big data analytics and efficient modeling in healthcare the following issues are take for this proposed research.

- To Identify patients at high-risk and to ensure they get the treatment they need, to develop algorithms to predict the number of days a patient will spend in a hospital.
- This research aims to accelerate the process of bringing innovations into practice through the linking of practitioners and researchers across the spectrum of biomedicine.
- To develop an appropriate model, it is necessary to compare and refine models derived from a diversity of cohorts, patient-specific features, and statistical frameworks.
- To establish new patient-stratification principles and for revealing unknown disease correlations.
- To learn a distance by multiple parties without data sharing and Interactive metric update and how to interactively update an existing distance measure.
- To extract of the important and relevant features and to extract the most relevant images for a given query.

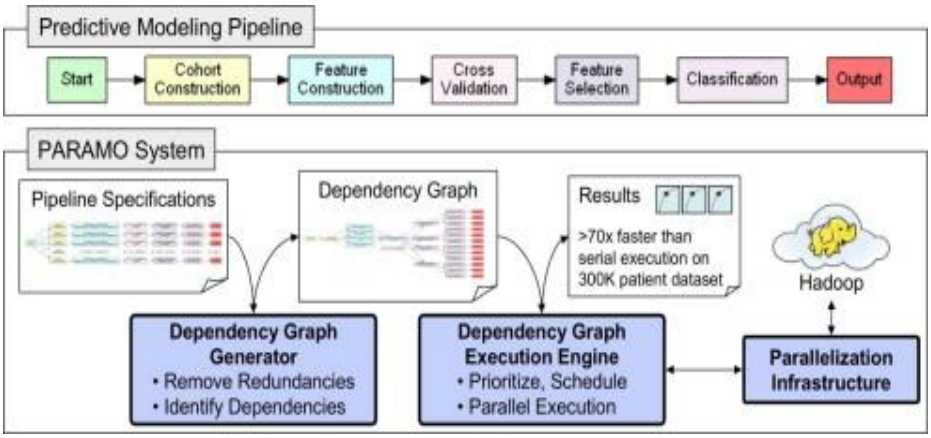
IV. METHODOLOGY

4.1 Health care providers can develop new strategies to care for patients before; it's too late and reduces the number of unnecessary hospitalizations. Improving the health of patients while decreasing the costs of care.

4.2 Biomedical informaticians interact with key stakeholders across the translational medicine spectrum (e.g., biologists, clinicians/clinical researchers, epidemiologists, and health. The success of translational medicine will depend not only on the addition of biomedical informaticians to translational medicine teams.

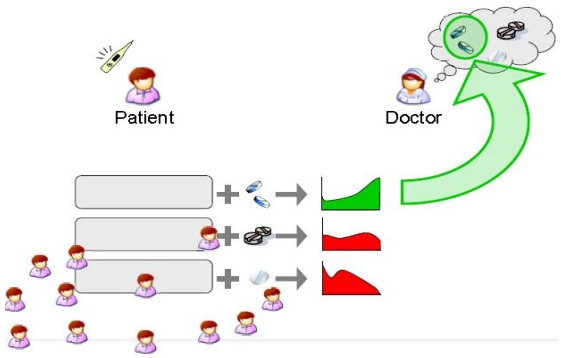


4.3 An efficient parallel predictive modeling platform can be developed for EHR data. This platform can facilitate large-scale modeling endeavors and speed-up the research workflow and reuse of health information. This platform is only a first step and provides the foundation for our ultimate goal of building analytic pipelines that are specialized for health data researchers.

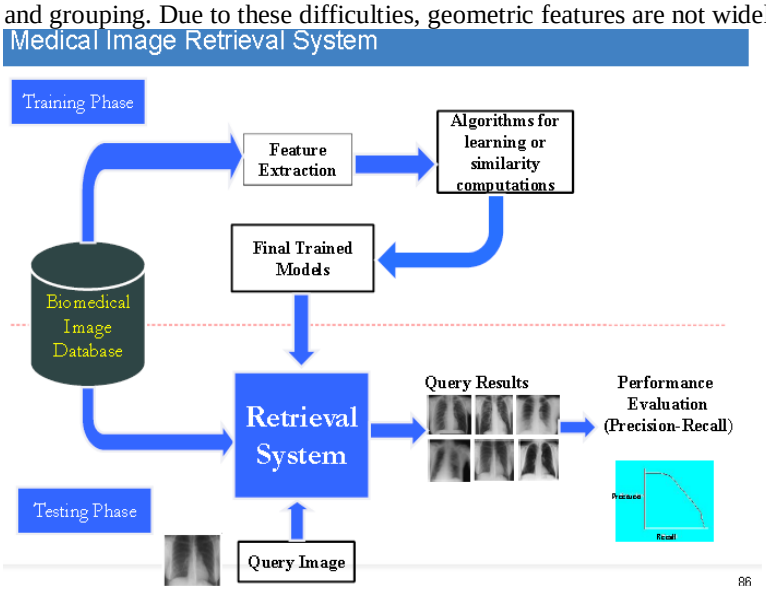


4.4 Integrating EHR data with genetic data will also give a finer understanding of genotype–phenotype relationships. However, a broad range of ethical, legal and technical reasons currently hinder the systematic deposition of these data in EHRs and their mining. Here, we consider the potential for furthering medical research and clinical care using EHR data and the challenges that must be overcome before this is a reality

4.5 Inter-patient similarity metrics can potentially help datamining researchers answer complex clinical questions for large populations. The issues with this task remain challenging. This study represents a formal attempt to address and discuss some of the underlying issues with inter-patient distance using ontology and information content principles as tools.



4.6 A suitable shape representation should be extracted from the pixel intensity information by region-of interest detection, segmentation, and grouping. Due to these difficulties, geometric features are not widely used.



V. CONCLUSION

There are enormous amount of data spread over across various healthcare domains. If these data and its sources are organized in predetermined and well defined way, it can be utilized to minimize the cost of research and to maximize the efficiency of healthcare knowledge. Big data analytics is a promising right direction which is in its infancy for the healthcare domain. Healthcare is a data-rich domain. As more and more data is being collected, there will be increasing demand for big data analytics. Unraveling the "Big Data" related complexities can provide many insights about making the right decisions at the right time for the patients. Efficiently utilizing the colossal healthcare data repositories can yield some immediate returns in terms of patient outcomes and lowering care costs. Data with more complexities keep evolving in healthcare thus leading to more opportunities for big data analytics.

REFERENCES

[1] M. Viceconti, P. Hunter and R. Hose, "Big Data, Big Knowledge: Big Data for Personalized Healthcare", *IEEE journal of biomedical and health informatics*, vol. 19, no. 4, pp. 1209-1215, Jul. 2015.

[2] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong and G. Z. Yang, "Big Data for Health", *IEEE journal of biomedical and health informatics*, vol. 19, no. 4, Jul. 2015.

[3] Yunchuan *et al*., "Internet of Things and Big Data Analytics for Smart and Connected Communities", *IEEE Access*, vol. 4, pp. 766-773, 2016.

[4] X. W. Chen and X. Lin, "Big Data Deep Learning: Challenges and Perspectives", *IEEE Access*, vol. 2, pp. 514-525, 2014.

[5] A. Fahad *et al*., "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis", *IEEE Transactions On Emerging Topics In Computing*, vol. 2, No. 3, pp. 267-279, 2014.

[6] L. Xu *et al*., "Information Security in Big Data: Privacy and Data Mining", *IEEE Access*, vol. 2, pp. 1149-1176, 2014.

[7] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard and K. Najarian "Big Data Analytics in Healthcare", *Hindawi Publishing Corporation, BioMed Research International*, vol. 2015, article ID 370194, 16 pages.

[8] V. Sujatha *et al*., "Bigdata analytics on Diabetic Retinopathy Study (DRS) on real-time data set identifying survival time and length of stay", *Procedia Computer Science*, vol, 87, pp. 227-232, 2016.

[9] S. Vennila and J. Priyadarshini., "Scalable Privacy Preservation in Big Data A Survey", *Procedia Computer Science*, vol. 50, pp. 369 – 373, 2015.

[10] Kovalchuk *et al*., "Personalized Clinical Decision Support with Complex Hospital-Level Modelling", *Procedia Computer Science*, vol. 66, pp. 392–401, 2015.

[11] J. Cunha, C. Silvaa and M. Antunes "Health Twitter Big Bata Management with Hadoop Framework", *Procedia Computer Science*, vol. 64, pp. 425 – 431, 2015.

[12] Z. Liu *et al*., "Quantitative Evaluation of Decision Effects in the Management of Emergency Department Problems", *Procedia Computer Science*, vol. 51, pp. 433 – 442, 2015.

[13] S. Ma and Y. Arjun, "Health Monitoring System by Prognotive Computing using Big Data Analytics", *Procedia Computer Science*, vol. 50, pp. 602 – 609, 2015.

[14] A. Abbas *et al*., "Personalized healthcare cloud services for disease risk assessment and wellness management using social media", *Pervasive and Mobile Computing*, vol. 28, pp. 81–99, Jun. 2016.

[15] F. Zhang *et al*., "A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications", *Future Generation Computer Systems*, vol. 43–44, pp.149–160, 2015.

[16] Y. Wang, L. Kung and T. A. Byrd, "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations", *Technological Forecasting and Social Change*, Feb. 2016.

[17] K. Kambatla *et al*., "Trends in big data analytics", *Journal of Parallel and Distributed Computing*, vol. 74, Issue 7, pp. 2561–2573, Jul. 2014.

[18] J. Wang, M. Qiu and B. Guo, "Enabling real-time information service on telehealth system over cloud-based big data platform", *Journal of Systems Architecture*, May 2016.

[19] S. M. DeJong, "The Future of Technology in Health Care" in *Blogs and Tweets, Texting and Friending Social Media and Online Professionalism in Health Care*, Chapter 12, pp. 151–163, 2014.

[20] P. Nadkarni, "Conclusions: The Learning Health System of the Future", *Clinical Research Computing, A Practitioner's Handbook*, Chapter 11, pp. 205–216, 2016.

[21] M. Legg, "Standardisation of test requesting and reporting for the electronic health record", *Clinica Chimica Acta*, vol. 432, pp. 148–156, May 2014, in Harmonization of Laboratory Testing - A global activity.

[22] A. T. Janke *et al*.,"Exploring the Potential of Predictive Analytics and Big Data in Emergency Care", *Annals of Emergency Medicine*, vol. 67, Issue 2, pp. 227–236, Feb. 2016.

[23] L. A. Winters-Miner *et al*., "The Predictive Potential of Connected Digital Health", Practical Predictive Analytics and Decisioning Systems for Medicine, Chapter 17, pp. 975–988, 2015.

[24] G. B. Melton *et al.*, "Inter-patient distance metrics using SNOMED CT defining relationships", Journal of Biomedical Informatics,Vol.39, Issue 6 ,pp.697-705, Dec. 2006.

[25] P.B.Jensen *et al*.,"Mining electronic health records:towards better research applications and clinical care", Nature Reviews Genetics 13,pp. 395-405 ,June 2012.

[26] Kenney Ng *et al., "PARAMO:* A PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records", Journal of Biomedical Informatics,Vol. 48,pp. 160-170, Apr. 2014.

[27] Jimeng Sun and C. K. Reddy., "Big Data Analytics For Healthcare",Tutorial presentation at the SIAM International Conference on Data Mining, Austin, TX, 2013.

