Data Analytics in Healthcare– Literature Survey

Team Members :

Nishanthi V - 913119104061
Keerthika R - 913119104044
Mathumitha M - 913119104053
Isha Parasu B - 913119104033

Abstract :

Finland

The current study performs a systematic literature review (SLR) to synthesise prior research on the applicability of big data analytics (BDA) in healthcare. The SLR examines the outcomes of 41 studies, and presents them in a comprehensive framework. The findings from this study suggest that applications of BDA in healthcare can be observed from five perspectives, namely, health awareness among the general public, interactions among stakeholders in the healthcare ecosystem, hospital management practices, treatment of specific medical conditions, and technology in healthcare service delivery. This SLR recommends actionable future research agendas for scholars and valuable implications for theory and practice.

Literature Survey on Data Analytics in Healthcare :

The concept of "big data" is not new; however the way it is defined is constantly changing. Various attempts at defining big data essentially characterize it as a collection of data elements whose size, speed, type, and/or complexity require one to seek, adopt, and invent new hardware and software mechanisms in order to successfully store, analyze, and visualize the data. Healthcare is a prime example of how the three Vs of data, velocity (speed of generation of data), variety, and volume, are an innate aspect of the data it produces. This data is spread among multiple healthcare systems, health insurers, researchers, government entities, and so forth. Furthermore, each of these data repositories is siloed and inherently incapable of providing a platform for global data transparency. To add to the three Vs, the veracity of healthcare data is also critical for its meaningful use towards developing translational research.

Despite the inherent complexities of healthcare data, there is potential and benefit in developing and implementing big data solutions within this realm. A report by McKinsey Global Institute suggests that if US healthcare were to use

big data creatively and effectively, the sector could create more than $300 billion in value every year. Two-thirds of the value would be in the form of reducing US healthcare expenditure. Historical approaches to medical research have generally focused on the investigation of disease states based on the changes in physiology in the form of a confined view of certain singular modality of data. Although this approach to understanding diseases is essential, research at this level mutes the variation and interconnectedness that define the true underlying medical mechanisms. After decades of technological laggard, the field of medicine has begun to acclimatize to today's digital data age. New technologies make it possible to capture vast amounts of information about each individual patient over a large timescale. However, despite the advent of medical electronics, the data captured and gathered from these patients has remained vastly underutilized and thus wasted.

Important physiological and pathophysiological phenomena are concurrently manifest as changes across multiple clinical streams. This results from strong coupling among different systems within the body (e.g., interactions between heart rate, respiration, and blood pressure) thereby producing potential markers for clinical assessment. Thus, understanding and predicting diseases require an aggregated approach where structured and unstructured data stemming from a myriad of clinical and nonclinical modalities are utilized for a more comprehensive perspective of the disease states. An aspect of healthcare research that has recently gained traction is in addressing some of the growing pains in introducing concepts of big data analytics to medicine. Researchers are studying the complex nature of healthcare data in terms of both characteristics of the data itself and the taxonomy of analytics that can be meaningfully performed on them.

In this paper, three areas of big data analytics in medicine are discussed. These three areas do not comprehensively reflect the application of big data analytics in medicine; instead they are intended to provide a perspective of broad, popular areas of research where the concepts of big data analytics are currently being applied.

Image Processing. Medical images are an important source of data frequently used for diagnosis, therapy assessment and planning. Computed tomography (CT), magnetic resonance imaging (MRI), X-ray, molecular imaging, ultrasound, photoacoustic imaging, fluoroscopy, positron emission tomography-computed tomography (PET-CT), and mammography are some of the examples of imaging techniques that are well established within clinical settings. Medical image data can range anywhere from a few megabytes for a single study (e.g., histology images) to hundreds of megabytes per study (e.g., thin-slice CT studies comprising upto 2500+ scans per study). Such data requires large

storage capacities if stored for long term. It also demands fast and accurate algorithms if any decision assisting automation were to be performed using the data. In addition, if other sources of data acquired for each patient are also utilized during the diagnoses, prognosis, and treatment processes, then the problem of providing cohesive storage and developing efficient methods capable of encapsulating the broad range of data becomes a challenge.

Signal Processing. Similar to medical images, medical signals also pose volume and velocity obstacles especially during continuous, high-resolution acquisition and storage from a multitude of monitors connected to each patient. However, in addition to the data size issues, physiological signals also pose complexity of a spatiotemporal nature. Analysis of physiological signals is often more meaningful when presented along with situational context awareness which needs to be embedded into the development of continuous monitoring and predictive systems to ensure its effectiveness and robustness.

Currently healthcare systems use numerous disparate and continuous monitoring devices that utilize singular physiological waveform data or discretized vital information to provide alert mechanisms in case of overt events. However, such uncompounded approaches towards development and implementation of alarm systems tend to be unreliable and their sheer numbers could cause “alarm fatigue” for both care givers and patients. In this setting, the ability to discover new medical knowledge is constrained by prior knowledge that has typically fallen short of maximally utilizing high-dimensional time series data. The reason that these alarm mechanisms tend to fail is primarily because these systems tend to rely on single sources of information while lacking context of the patients’ true physiological conditions from a broader and more comprehensive viewpoint. Therefore, there is a need to develop improved and more comprehensive approaches towards studying interactions and correlations among multimodal clinical time series data. This is important because studies continue to show that humans are poor in reasoning about changes affecting more than two signals.

Genomics. The cost to sequence the human genome (encompassing 30,000 to 35,000 genes) is rapidly decreasing with the development of high-throughput sequencing technology. With implications for current public health policies and delivery of care, analyzing genome-scale data for developing actionable recommendations in a timely manner is a significant challenge to the field of computational biology. Cost and time to deliver recommendations are crucial in a clinical setting. Initiatives tackling this complex problem include tracking of 100,000 subjects over 20 to 30 years using the predictive, preventive, participatory, and personalized health, refered to as P4, medicine paradigm as

well as an integrative personal omics profile. The P4 initiative is using a system approach for

(i) analyzing genome-scale datasets to determine disease states,
(ii) moving towards blood based diagnostic tools for continuous monitoring of a subject,
(iii) exploring new approaches to drug target discovery, developing tools to deal with big data challenges of capturing, validating, storing, mining, integrating, and finally

modeling data for each individual. The integrative personal omics profile (iPOP) combines physiological monitoring and multiple high-throughput methods for genome sequencing to generate a detailed health and disease states of a subject. Ultimately, realizing actionable recommendations at the clinical level remains a grand challenge for this field. Utilizing such high density data for exploration, discovery, and clinical translation demands novel big data approaches and analytics.

Despite the enormous expenditure consumed by the current healthcare systems, clinical outcomes remain suboptimal, particularly in the USA, where 96 people per 100,000 die annually from conditions considered treatable. A key factor attributed to such inefficiencies is the inability to effectively gather, share, and use information in a more comprehensive manner within the healthcare systems. This is an opportunity for big data analytics to play a more significant role in aiding the exploration and discovery process, improving the delivery of care, helping to design and plan healthcare policy, providing a means for comprehensively measuring, and evaluating the complicated and convoluted healthcare data. More importantly, adoption of insights gained from big data analytics has the potential to save lives, improve care delivery, expand access to healthcare, align payment with performance, and help curb the vexing growth of healthcare costs.