J Supercomput (2020) 76:1754–1799
https://doi.org/10.1007/s11227-017-2222-4

Big data analytics enhanced healthcare systems: a review

Sarah Shafqat[1] · Saira Kishwer[1] · Raihan Ur Rasool[2] · Junaid Qadir[3] · Tehmina Amjad[1] · Hafiz Farooq Ahmad[4]

Published online: 3 February 2018

Abstract There is increased interest in deploying big data technology in the healthcare industry to manage massive collections of heterogeneous health datasets such as electronic health records and sensor data, which are increasing in volume and variety due to the commoditization of digital devices such as mobile phones and wireless sensors. The modern healthcare system requires an overhaul of traditional healthcare software/hardware paradigms, which are ill-equipped to cope with the volume and diversity of the modern health data and must be augmented with new "big data" computing and analysis capabilities. For researchers, there is an opportunity in healthcare data analytics to study this vast amount of data, find patterns and trends within data

✉ Sarah Shafqat
sarah.shafqat@gmail.com

Saira Kishwer
saira.abbasi11@gmail.com

Raihan Ur Rasool
raihan.rasool@live.vu.edu.au

Junaid Qadir
junaid.qadir@itu.edu.pk

Tehmina Amjad
tehminaamjad@iiu.edu.pk

Hafiz Farooq Ahmad
hfahmad@kfu.edu.sa

1 International Islamic University, IIUI, Islamabad, Pakistan

2 Victoria University, Melbourne, Australia

3 Information Technology University (ITU), Lahore, Pakistan

4 College of Computer Sciences and Information Technology (CCSIT), King Faisal University, Alahssa, Saudi Arabia

and provide a solution for improving healthcare, thereby reducing costs, democratizing health access, and saving valuable human lives. In this paper, we present a comprehensive survey of different big data analytics integrated healthcare systems and describe the various applicable healthcare data analytics algorithms, techniques, and tools that may be deployed in wireless, cloud, Internet of Things settings. Finally, the contribution is given in formation of a convergence point of all these platforms in form of SmartHealth that could result in contributing to unified standard learning healthcare system for future.

Keywords Healthcare analytics · Big data · Cloud computing · Knowledge management · Learning healthcare system

1 Introduction

Healthcare industry generates abundant amount data that revolves around patients, drugs, diseases, cures, research, and many more [1]. Trends have been identified to digitize all this data for patient care using healthcare analytics by record keeping, compliance and regulatory requirements [2]. The healthcare big data involves all the clinical data from Computerized Physician Order Entry (CPOE) and clinical decision support systems—physicians compiled reports, prescriptions, medical imaging, laboratory, pharmacy, insurance and other administrative data; electronic patient records (EPRs); machine generated/sensor data, from monitoring vital signs; social media posts including Twitter feeds, blogs, Web sites, Facebook updates and other platforms; and minimal patient care data including emergency care data, news feeds, and medical journals. The big data storage for healthcare promises to improve the quality of healthcare while reducing the cost at the same time. It has potential to support various medical and healthcare functions inclusive of—clinical decision support, disease surveillance, and population health management. According to reports health data residing only in the USA goes beyond 150 exabytes in 2011 and it has the capacity to go further and beyond to zettabytes (10^{21} bytes) and yottabytes (10^{24} gigabytes). The analysis of big data itself becomes the bottleneck because of its massiveness. Therefore, healthcare domain experts are looking toward computer sciences to investigate and come up with the solutions to transform data into information and knowledge. There are emerging technologies in the field of data science such as Hadoop, unsupervised learning (finding hidden patterns in data), graph analytics, and natural language processing (where knowledge is extracted from documents enabling computers to understand the textual language of humans). The big heterogeneous data in the medical field are now overwhelming the intuitive abilities of the healthcare practitioners. The need of algorithms is highly recognized through which correlations are developed between all the associated factors and features. Thus, a massive opportunity lies ahead promoting medicine as an information science laying the foundation for learning healthcare system (LHS) [3]. Professor Friedman of Institute of Medicine (IoM) defined the cycle of processes in LHS. LHS vision is to acknowledge any healthcare delivery system that is worked upon at institutional, national, or international level as its by-product [4].

Conceptually, the architectural framework for big data healthcare analytics is not unlike the traditional healthcare informatics architecture [1]. The key difference is apparent when we consider how processing is being done. Traditionally, healthcare projects involved business intelligence tool deployed on a standalone system. When the 'big data' is involved, processing requires to be divided into multiple nodes for execution, bringing distributed processing [5] in use. Healthcare networks need redefining [6] as data comes from internal as well as external sources residing at multiple locations. This effort is demonstrated in [7] as IoT healthcare network (IoThNet). Sources and data types include [1]:

- Data from social media platforms as clickstream, and interactive data from Facebook, Twitter [8], LinkedIn, blogs, smart apps, and Web sites with a health plan, etc.
- Machine to machine data: sensor readings, meters, and other vital sign devices.
- Big transaction data: records for healthcare claims [9] and billings, available in semi-structured and unstructured formats.
- Biometric data: fingerprints, genetics, scans of retina, handwriting, blood pressure, and other diagnostic reports and data.
- Human generated data: semi-structured data, as well as unstructured data such as electronic medical records (EMRs), physicians' notes, email, and paper documents.

In Sect. 5, this paper presents the analysis of various healthcare platforms/frameworks embedded with analytics algorithms and techniques that have been used till now for prevention, prediction, diagnostic, and treatment of various chronic diseases like cancer, heart disease, diabetes and kidney disease. There are various frameworks and systems that have been proposed and being developed to contribute toward healthcare domain. Through a detailed taxonomy in Table 2, this paper highlights some of the key frameworks and healthcare systems analyzing their contribution and achievements as well as hindrance faced. Previous surveys and reviews that have been done were limited in scope as they focused on algorithms [10–20] or a framework [7,8,21–28] but did not touch both aspects of healthcare technology. This paper rigorously explores various healthcare analytics platforms to find the convergence point for the foundation of SmartHealth [29] to strengthen LHS [3,4,28] vision. The paper clearly highlights the effectiveness and importance of following standard healthcare conventions and international standard bodies like Health Insurance Portability and Accountability Act (HIPAA) [30] and World Health Organization (WHO) [7,31].

The rest of this paper is organized in the following way (Fig. 1). Section 2 describes various healthcare related conventions and terminologies that need to be understood by healthcare analytics. Section 3 focuses on various frameworks and systems known for healthcare analytics. Section 4 is devoted to healthcare analytics tools and techniques. Section 5 further performs a high-level analysis through comparison of healthcare systems with underlying known platforms and technologies (Table 2). Furthermore, these frameworks and data analytics platforms have been thoroughly evaluated and discussed to find relationships between various data types or variables. Section 6 is devoted to limitations and strengths. While Sect. 7 defines the accuracy model that could be used for validation of healthcare analytics in future.

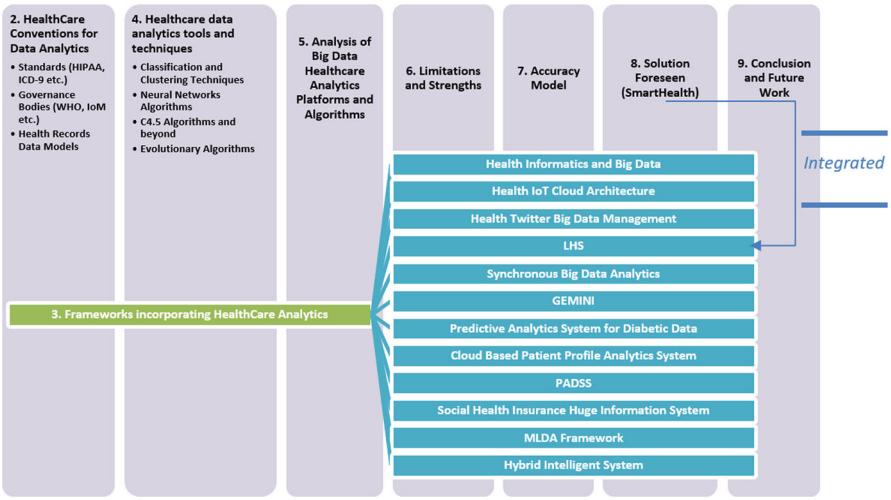

Fig. 1 Article structure

Section 8 proposes the solution foreseen and Sect. 9 concludes and reflect future vision.

2 Healthcare related conventions for data analytics

To apply healthcare data analytics, there is need to understand the standard conventions used in the healthcare industry and its organizational or data model [21]. The two terms are often used in normal routine clinical assessment:

Electronic Medical Record (EMR) is a computerized patient record of medical history in the provider organization. This record can incorporate data from referred specialists or pharmacists and laboratories but mostly it is valid within the provider organization and is used to avail services of that specific organization.

Electronic Health Record (EHR) is a higher level of EMR. An EHR [20] is composed of universal patient ID linked to lifetime medical history that is valid and may be shared across multiple provider organizations. The vision set by Health and Human Services (HHS) for a national EHR data store is to authorize patient the right to his medical history which he can share with other healthcare providers.

Therefore, provider centric EMRs are fed into patient centric EHR and the real outcome would be felt when EMRs would be integrated with practice management systems and EHRs are envisioned by research systems to support evidence-based healthcare delivery over the social or community-based cloud.

These recent few years have seen a tremendous adoption and value of EMR as given below:

Government endorsement in July 2004, the US department of Health and Human Services (HHS) incorporated a 10-year plan for establishing National Health Information Infrastructure (NHII), having an EHR for every American and a new network

to link health records nationwide. According to this plan, there comes an initiative for providing a quantum leap in patient power, doctor power, and effective healthcare.

Standardization [32] *of health record models* HHS commissioned IoM to design a standard EHR model as seen in an effort by [27] for healthcare industry using an ANSI standard [33] health level 7 (HL7) [30]. This initiative helped the formation of collaborative EHR, a broad-based consortium of private and public-sector healthcare organizations generating data models [21] for IoM approval. National EHR standard, as part of NHII, would help improve effective data sharing among all stakeholders. *Standard language* [24] uniformity is required for keeping a global standard that is understood by healthcare data analytics and information systems. This goal got a huge boost when the National Institute of Health National Library of Medicine had a five-year contractual agreement with the College of American Pathologists for giving license of SNOMEDCT (Systematized Nomenclature of Medicine—Clinical Terms). This contract enables universal access to machine-readable, clinically rich lexicon for standardized coding [34].

3 Frameworks incorporating healthcare analytics

The paper explores several frameworks falling in the domain of healthcare informatics and big data analytics. The summarized taxonomy of these platforms is shown in Table 2 in Sect. 5 mentioning all tools and techniques that were used for comparison.

3.1 Healthcare informatics and big data

There are considerably standardized [5,30] developments in healthcare informatics [35,36] with respect to big data being generated from a diverse range of data sources keeping security under consideration [6,37,38]. Data sources [21] are not just limited to clinical diagnosis or personalized medicine but includes imaging, genomics, metabolomics, proteomics and long-term psychological sensing of an individual. The knowledge being converged from big data is being tested for the new hypothesis in disease management for diagnosis, prevention, and providing personalized treatments. The last has revolutionized the healthcare [36] taking people out from isolation to the interconnected world of technology [7,8] and cellular network where healthcare is easily reachable for people.

There are many challenges needed to be addressed [35]. The massiveness of big data in healthcare [35] itself is a big issue. Considering the US healthcare system only five years ago gives us an estimation of data reaching 150 Exabyte (10^{18}). It would not take long that we would have to deal with data amounting to zettabytes (10^{21}) and yottabytes (10^{24}) when other countries also get into the picture. Data are coming from many platforms [7] that are mainly social media [8,36], real-time imaging, high throughput sequencing platforms, the point of care devices, mobile health technologies and wearable computing [39]. Big data has six Vs that also apply to healthcare data: value, volume, velocity, variety, veracity, and variability. Size is the foremost challenge as other challenges of heterogeneity and variety in data that is

coming in with speed has already gained attention in [20,23]. Data heterogeneity and variety result from structured and unstructured data entering from various platforms that are quantitative (e.g., laboratory tests, images, sensor data and gene arrays) and qualitative (e.g., demographics and free text) [23]. The veracity of data that is its trustworthiness is very crucial when data are coming from unmanaged sources like social media it becomes challenging. To reap the full benefit of big data in healthcare, it is important that it is analyzed in a coherent manner. Other challenges are in terms of social and legal technicalities comprising of data ownership, privacy, data stewardship, identification and governance [21].

An opportunity lies in integrating traditional health informatics with mobile health and social health [35,36]. Mobile and social health platforms are there to connect patients with doctors outside the clinical premises over the cloud [28]. In social health, the communication is expanded from between patient and doctor to patient to patient as well. Now, patients with acute and chronic diseases like diabetes, cancer, and heart disease are communicating over the social networks to discuss their experiences with each other. There are public health surveillance systems in place that correlate the possible emergence of asthma-related attacks in a polluted environment. Mobile messaging [24] is also being considered as a preventive therapeutic treatment for patient behavioral changes or in case of diabetes enforcing a possible change in lifestyle for positive results.

3.2 Health IoT cloud architecture

Internet of Things (IoT) has revolutionized the healthcare with promising the technological invasion, economy, and social prospects. State-of-the-art network architectures, systems, and industrial breakthrough have been studied to get an understanding of how IoT is evolving the healthcare technologies. Many IoT and eHealth policies and regulations are under review to find how the innovations in big data, ambient intelligence, and wearables are leveraging healthcare by proposed intelligent collaborative security model [7] and would facilitate the economy and societies to achieve sustainable development.

IoT [7] is a term given to connected nodes—whether these nodes represent people, places, machines, or anything else. Its advantages include automation of smart cities, traffic congestion, structural health, waste management, logistics, emergencies, security, and retail.

IoT healthcare network (IoThNet) supports access to IoT backbone, enabling healthcare tailored communication with the facilitation of transmission and reception of medical data. IoT healthcare network consists of topology, architecture, and platform being discussed here [7]. Topology is inclusive of physical configurations, application scenarios, activities, and use cases. The architecture consists of software organization of the system and hierarchical model. This platform has all the libraries in a framework within given environment.

IoT healthcare network [7] has become these days such that human body is connected to various sensors such as bio-patch, biochip and IMedPack those are recording each symptom and body condition which is being read through WBAN devices,

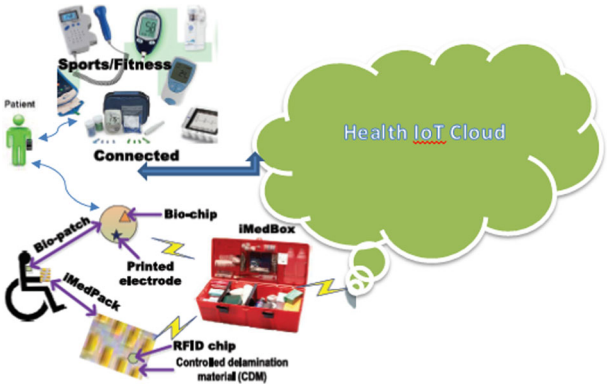

Fig. 2 Health IoT cloud architecture [7]

recorded and deployed [28] in data center of Health IoT cloud (Fig. 2). The health IoT cloud is conceptualized to receive deciphered information from various sources including but not limited to mobile apps like sports and fitness monitors, RFID bio sensors like iMedPack having bio-patch, social networks as modeled by [40,41] and adopted by WHO [42].

Big data service is required to process massive streams [36] of data coming in from IoT and needs a thorough understanding of data distribution while handling all the continued input load. An example is seen where healthcare sports data is being analyzed by an application to reach some conclusions. Smart healthcare apps are taking up that are equipped with sensors. Applications like Endomondo today track all daily activities that range from sports, diet routine, sleep habits over the social networks. Endomondo is a famous sports-oriented application catering more than 30 million users from all over the world. It is integrated with services like GPS and accelerometers that track devices along with their route for speed, distance, time and duration. Users can share their progress via social networks like Facebook and Twitter for comments and suggestions. A dataset up to 15,090 users was thus taken from Endomondo web server for analysis. In the study, there was data for 333,689 workouts that were obtained over the period of 5 months in 2014. The log file contained user profile, GPS trace, and workout summary. The sports tracker dataset summary is presented in Table 3 [36].

The data were analyzed with respect to various parameters like it showed that three continents having most users involved in sports are Asia, Europe, and America. The most played sport is running, cycling, and walking followed by others [36].

3.3 Health Twitter big data management

With advancement in social networks, there is a rapidly increasing data explosion that needs to be managed. Apache Hadoop with Mahout is a big data analytical tool not only employed by large private enterprises and businesses like Facebook and Google but its application is seen in healthcare institutions as well [8]. A generic functional architecture for big data analytics [8] is therefore proposed that can be used in different

scenarios [23,36] as it is used on Twitter-based health data. It is developed on Hadoop framework using Mahout. It gives us a picture of how health social data can give useful insights for a common user as well as practitioners and doctors.

The Twitter API with the twitter4j library is integrated with a web application using NoSQL database in MongoDB getting a stream of Twitter real health data as an input. This proposed framework [8] integrated with healthcare-specific systems can deliver an improved result in classifying both social health data and healthcare system specific data. In this framework, some specific keywords related to healthcare were selected to classify the health Twitter data. Twitter data were extracted at intervals of 30 minutes not to get blocked. The data being retrieved are unstructured tweets; therefore, MongoDB having NoSQL standard is used. Hadoop process the data using MapReduce algorithm to structure the data and represent it. HDFS stores this data in form of clustered chunks. The MapReduce functions are applied to further organize the key value data block for processing and treatment. After preprocessing of data, the learning and classification algorithms are applied by Hadoop Mahout for final results [8]. The resultant data are stored in a relational SQL that is available to any web service for use in web application. The web application analyses these results on basis of input hashtags.

The statistics and results [8] can be acquired to be used for various decision support systems for diagnosis and treatments of different diseases, habit changing, and disease prevention, etc. Certainly, it has clear benefits for future health care for the betterment of human life.

3.4 Learning healthcare system

In healthcare when laying the foundations for learning healthcare system [3,34], decisions regarding disease diagnostics through classification is done by identifying patterns of clinical practices taken by physicians. It is not easy to comply with the needs of stakeholder in healthcare whether they are patients, medical staff, administration and other governing bodies. It is already late for the adoption of big data approaches that has left the medical field not fully prepared for its aspiration of future precision medicine [3] to be designed for individualized healthcare using personalized information through learning. The systemized and accurate view of information about patients to clinicians is not apparent and the knowledge of risks and benefits is often vague. Even if the evidence exists for a decision in a specific case it is not always applicable to the patient. Predictions for personalized prognosis and response to treatment is required for improvement in informed decisions. Relationships between all factors including risks associated with drugs and devices, effective ways of prevention, diagnostics, prescriptions and related treatments need to be apparent and known for an understanding of a patient's case in a healthcare system that is an integral part of society.

Developing clusters of patient groups is one way to form taxonomies of diseases on basis of their similarity because of shared characteristics and outcomes. Empirical classification [3] is used for selection of best treatment strategy with predicted results. The knowledge of best possible treatments is precedent in understanding the under-

lying mechanism of disease and its response to treatment. This is a learning approach to reproduce consistent mechanisms that would work efficiently in diversified settings and patients. Inductive reasoning and pattern recognition [3] in contrast to deductive reasoning is based on learning through observations to evolve conceptual models and tools for informed decisions. The approach is validated through testing the consistency of results and conclusions.

Inductive reasoning [3] is less certain to identify causes and more relates to forming the confidence level than to reach definitive conclusions. For example, there is no such experiment to validate that smoking causes cancer but a high confidence level is gained through observation. The criterion is required that would assist researchers to interpret results of millions and trillions of observations for rapid decision making. While retaining the complexity of patients and medical decision making big data approaches highlights the interactions between all the factors underlying to reach future of precision medicine diminishing the effects of associated risks and outcomes. Researchers are good at finding patterns without knowledge of the outcome.

3.5 Synchronous big data analytics

Remote clinics are found to be a better therapeutic treatment for elders at the convenience of their homes. Telehealth services [25] are a good remedy with the introduction of gigabit networking through Google fiber installed at smart homes in some of US cities like Kansas. It has the capabilities of sensing, high-definition video-based communications and cloud computing. Telehealth system enabled in smart city saves cost with fewer hospital visits for elder care.

Thus, a PhysicalTherapy-as-a-Service (PTaaS) [25] is introduced for elderly in telehealth connecting remote physical therapist with an elderly at their home. It is in form of interactive interface [25] with low tenancy network connectivity developed on top of Microsoft Kinect having motion sensors. The platform used for development is open source Kinect API, C# language within Windows Presentation Foundation (WPF) technologies integrated into Dot Net Framework 4.5. Hardware components that are used apart from sensors are network monitoring services and measurement point appliances by Nerada Metrics [25] for the application. This application is integrated with cloud storage provided by Global Environment for Network Innovations and applied on it are synchronous analytics for data streams generated in PTaaS sessions.

This synchronous big data nature of PTaaS was investigated [25] by conducting various experiments and in a real-time scenario based in homes for elderly in Kansas City connected by Google fiber and clinic that was situated in Columbia, Missouri. The results clearly demonstrated the challenges for network configurations and time synchronization in online analytics applied on PTaaS data streams. This is a basic limitation of using synchronous big data analytics. The physical therapeutic activities monitored were an analysis of walking patterns, balance, and sway detection.

The PTaaS is compared [25] to the ideal network health scenarios [35,36] and the worse cases. PTaaS was also evaluated for quality based on the comparison with face-to-face therapeutic sessions by getting users feedback and judging on the effectiveness criteria.

3.6 GEMINI: an integrative healthcare analytics system

GEMINI allows a real-time point of care health services being provided to patients interacting with the expert doctors and staff. It is known as an Integrative Healthcare analytics system incorporating various algorithms for feature selection, clustering, classification, and prediction techniques. GEMINI [24,26] has two modules in it that are profiling and analytics.

The profiling [24] saves patients' profiles collected from various data sources in form of patient profile graph. The data are structured as well as unstructured. Structured data include demographics like age and gender information, and laboratory results. Unstructured data includes doctors' notes that is free text. The analytics module [24,26] is used to analyze patient profile graphs to give some useful results in form of predictions. For example, it can be inferred that patient's diabetes mellitus is not well-controlled. To understand free text, there are various Natural Language Processing (NLP) [34] engines like MedLEE and cTAKES. and also, some medical dictionaries to understand medical conventions like Unified Medical Language System (UMLS) [24].

GEMINI gives a self-learning knowledge base [24] keeping all experts in a feedback loop to gather, infer or ascertain various results and conclusions. This knowledge base is capable of forming domain specific relationships [24] between two or more concepts HbA1c test is used as a measure to monitor diabetes mellitus (DM) control by fully utilizing the power of semantic computing. GEMINI is used to predict in different scenarios. One case where it identifies patients with probable high risk of heart disease in near future or in another case where it predicts the probability of patients to be readmitted within 30 days, etc.

3.7 Predictive analytics system for diabetic data

DM is one of the uncommunicable disease and major health hazard in developing countries including Pakistan. Predictive methodology for analyzing diabetes patients' data is thus devised to predict types of diabetes prevalent and complications associated with it highlighting the possible treatment. Hadoop using MapReduce is the chosen platform to process collected data in data warehouse. The architecture of predictive analytics system is given in Figure 1 in [61]. For coming up with the possible treatment in diabetes, it requires certain parameter values such as glucose level, blood pressure, serum insulin, body mass index (BMI), diabetes pedigree, and patient profile information like age and previous pregnancies.

To find patterns for prediction the analytics system must go through [27]: (i) association rule mining, (ii) clustering, (iii) classification, (iv) statistics for accuracy and (v) applying predefined deductive rules on data.

MapReduce in Hadoop maps the large dataset splitting into smaller chunks and assigning to worker nodes. Worker nodes perform the pattern matching task with help of data nodes and stores the processed data in intermediate disks. When the query from client is received at master node the worker node performs the reduce task at the intermediate disk and results retrieved would be sent to server end [61].

3.8 Cloud-based patient profile analytics system

A cloud-based patient profile analytics system [26] is devised to cater diabetic patients' data. Patient Profile Graph is there to store this data and to represent it in holistic view. The required information is inferred for clinical and administration purposes while performing predictive analytics as well. It is to predict unplanned readmissions, and chances of getting diabetes by observing family history and dietary plan. Data analytics [24] integrated in cloud computing enables mainstream data processing, storage and distribution. The challenge lies in moving big data in and out of the cloud for health institutions and organizations that is in terabytes and petabytes at the speed that is needed.

The proposed system keeps the healthcare specialists in loop for feedback to keep updating the self-learning knowledge base [26]. The profiling and analytics modules are inbuilt in the proposed system to store patient profiles and perform analytics [26] over it. Analytics system identifies the concepts and relationships in patient profiles using feature selection techniques that would form the basis of analyzing to reach conclusions. Feature selection is done by input from doctors or is automated. Various analytics algorithms and techniques are applied to these features and training data. To accurately label the inferred data training, data are grouped with similar patients and doctors are asked for suggestions. After having training data labeled the analytics is applied using various machine learning algorithms and picks up the rows that show low confidence on applying labels and get the doctors to label.

3.9 Cloud-based healthcare platform

Disease outbreaks are happening all over and always. Computational prediction, identification, confirmation, and responsiveness to these diseases is important as well. Therefore, Predictive Analytical Decision Support System (PADSS) [22] integrates in a cloud-based healthcare platform that is Message Oriented Middleware (MOM). It connects healthcare organizations to share data using a customized Health Level Seven (HL7) [30] platform having Fast Healthcare Interoperability Resources (FHIR) [33] specification to provide a proper structural format to exchange data for prediction of disease outbreaks in real-time scenario. HL7 is a nonprofit international body for the development of interoperability standards [33] that lays down specifications for healthcare software applications. FHIR in HL7 enables clinical and administrative data exchange on international standards between healthcare applications. FHIR data model is extensible to allow applications to modify on need basis while using a set of customized data structures and resources. FHIR uses XML and JSON-based formats to simplify system level communication using common library of interfaces. PADSS [22] is built on top of PREVENT that used statistical report to predict outbreak of a disease in real-time style where PADSS is designed to extract and load data from messages received from various healthcare organizations finding patterns in historical and transactional data aiming to predict disease outbreaks. Google cloud data store is used to provide schema less NoSQL scalable persistence services at PREVENT persistence layer. PREVENT [22] was upgraded to perform real-time large big

data analysis jointly with Complex Event Processing (CEP) that would resemble to online analytical processing (OLAP). To comply with the upgradation, PREVENT was modified to include Google Big Query. The patient data hold values like patient diagnostics, location, and incidence rates. This information leads to identifying patterns that could match to historical data to predict or identify a disease occurrence early. PADSS middleware application was validated and tested [22] by running two simulations being deployed on Google Cloud Platform to access its responsiveness, scalability, and reliability.

It confirmed the improvement in rate of prediction and analysis of disease occurrence for more precise decision making [21]. It was recognized that tuning up of SQL queries would lead to better results in time constrained scenarios. When dealing with large heterogeneous data, it is better to setup an upper limit for the number of delimited risk zone points in the map for filtration of results. Lastly, to achieve quality in results the profile settings need to be done for use of adaptive rules for effective decision making.

3.10 A proposed framework applying big data analytics on social insurance huge information of patients

There is a huge data flood with crude patients' health information in form of XLS sheets in the social insurance sector [9] that is analyzed to perform three types of information quality administration systems check: (i) information gushing, (ii) information order, and (iii) group examination module. There are mainly two group modules. Quality bunches and esteem groups jointly work to guarantee the worth of information esteem connecting patients' data, e.g., age, gender, and sugar level. The worth groups would be responsible to find out if there is any missing detail in patient's data for a social insurance case through a probabilistic relapse examination. On top of quality bunches and esteem groups, there is a supervisor group module that oversee using parallel processing worldview.

The proposed framework having this module performs a practical and proficient asset designation administration as per patient's data [9]. This proposed system has defined a way to analyze big data apart from traditional way using Hadoop framework. This model is adaptable to be used for other novel data mining techniques for analyzing healthcare big data to investigate, restorative examination and clinical decision support for a specific patient.

3.11 Multi-Level Data Analysis (MLDA) framework

Considering the case study in previous research for analyzing healthcare domain, where Multi-Level Data Analysis (MLDA) framework [23] is exploited over to analyze the treatments of diabetic patients (Fig. 3).

In the healthcare domain, previous studies have shown that there are many correlated diseases affecting groups of patients—e.g., heart valve disease has been diagnosed using cluster methods [43]. In another example [44], breast cancer has also been diagnosed using cluster techniques finding patterns for benign and malignant tumors

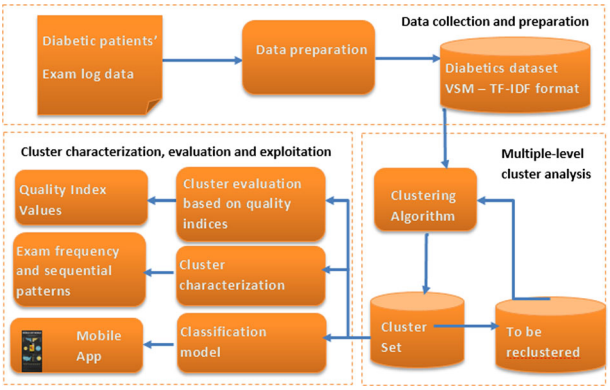

Fig. 3 MLDA framework on treatments of diabetic patients [23]

within tutor features. K-means algorithm is also used to cluster a set of patient records aiming to recognize various features of patients affected by heart disease [45].

There has been considerable research on exploiting clustering techniques on data of diabetic patients [20,44,45]. There have been many issues identified as gait patterns [46], food analysis [47] relate diabetes to various associated risk factors [43,44], finding similar medical treatments [45–48,62] and analyzing several imputation techniques [48,49]. Antonelli et al. [48] especially focuses on using K-means algorithm for analyzing various imputation techniques based on diabetic patients' datasets. In our framework, a relatively different approach has been used and patients with similar case histories are grouped together and classified into a set of predefined classes. Each cluster was detailed with sequential patterns for finding out the way examinations were included and spanned over time.

There is also a considerable research going on linking the medical care with mobile technologies [50–52] as with cloud [28] in [7,22] and devices as the capability and its wide use has drawn interest toward providing various useful services for user-generated data [23]. [45,47,48] revolves around developing a diagnostic app for diabetic patients that provides distributed end-to-end pervasive healthcare system on mobile using neural network computations. Similarly, K-means was used [45,47,48] as an unsupervised classification technique for automatically detecting seizures [52] through mobile-based approach. Least Square Support Vector and Generalized Discriminant Analysis machine models are also used for diagnosing patients with diabetes [50]. In this framework, a two-tier architecture approach is used where ubiquitous patient classification is achieved through mobile app labeling unlabeled examination histories and using multiple level cluster analysis as adopted in [23]. This knowledge is efficiently and effectively made available to various user profiles whether they are patients or medical staff.

3.12 Hybrid Intelligent System

Finally, the synergy is produced in [16–19,43,44,49] which are mostly family of genetic algorithms [18] that are evolving. In recent analytics systems and models, we

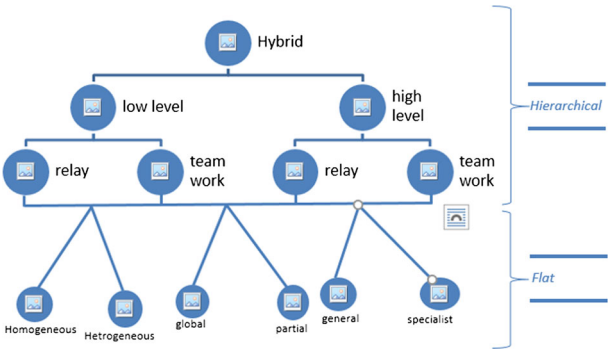

Fig. 4 Hybrid metaheuristic for evolutionary algorithms [16]

witnessed that combination of algorithms was used to get an optimum solution. This combination of AI algorithms is often known as Hybrid Intelligent System [16] (Fig. 4) to make common sense, extract knowledge, think like humans, dealing with uncertainty and imprecision, and adapt to rapidly changing and unknown environment. Similarly, it is also applied to diverse clinical scenarios, like breast cancer diagnosis, diagnosis of coronary artery stenosis, analysis of microcalcification on digital mammograms, control of depth of anesthesia, and assessment of myocardial viability.

When devising a Hybrid Intelligent System or model there are various methodologies to combine different algorithms as depicted in Fig. 4. Low-level hybridization refers to single optimization (that is finding the best solution through the best available means) where one function of metaheuristic is replaced by another metaheuristic in the process of finding the best solution. High-level hybridization is self-contained and is not directly involved in internal workings of metaheuristics. The relay approach is another way to combine selected algorithms where the algorithms are applied one after another as a pipeline. Teamwork refers to parallel cooperating agents searching in a given solution space and saves processing time.

Four classes of frameworks [16] are generated from this: (i) low-level relay hybrid, (ii) low-level teamwork hybrid, (iii) high-level relay hybrid, and (iv) high-level teamwork hybrid. The hybrid algorithm goes flat it may become either homogeneous, heterogeneous, global, partial, general, or specialist. In homogeneous, the algorithms may be initialized with different initial solutions. In case of heterogeneous, best solutions (that are more near to estimated or expected solution) from the precedent iterations form the basis of initiating next algorithms and thus become adaptive. In a global hybrid, the research space is searched thoroughly. In a partial hybrid, the problem is divided into subproblems and each algorithm is given a search space of one of these subproblems. These subproblems communicate their optimal solutions to the next subproblem in pipeline keeping up with constraints and find an optimal solution as a whole. In general hybrid, the target problem remains the same where in specialist hybrid algorithms are combined to solve different problems. The optimized hybrid framework is chosen based on the problem complexity and need.

The concept of on-cloud healthcare clinic [53] is well established keeping in view the various evolutionary algorithms in synchronous [25] and asynchronous [17,19] mode, hybrid intelligent framework given in [16] and other healthcare analytics

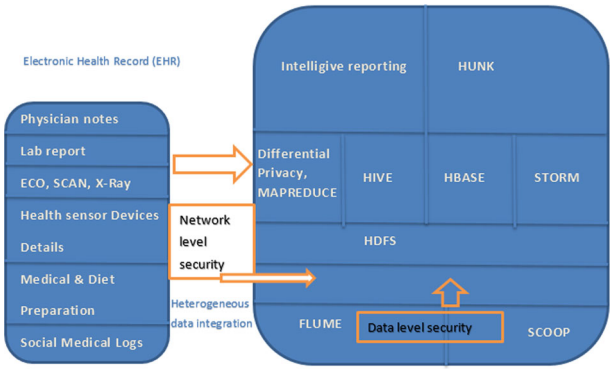

Fig. 5 Secured big data analytics architecture used in HealthCare industry [21]

systems viewed in [16,43,44,49] comprise as combination of algorithms, various healthcare analytics tools and techniques. Therefore, this study would best serve the purpose to evaluate which of these would best comply with our proposed healthcare investigative system, i.e., SmartHealth [29], in future. There has been a lot of work already done in e-healthcare domain shown in Table 3 of [53].

4 Healthcare data analytics tools and techniques

The healthcare domain is information rich but requires transforming it into knowledge. The effective tools and techniques are formulated to discover hidden relationships and trends in data. Kavakiotis et al. [54] refers to a detailed systematic review on application of data mining and machine learning techniques to perform analytics on diabetes mellitus (DM) for prediction and diagnosis, finding complications associated with it and its linkage with the genetic background and environment to assist betterment in healthcare management. There are various data mining techniques being used in the healthcare industry for various diseases using classification method that is demonstrated through successful case studies, these are mainly rule-based, artificial neural networks and decision tree [10].

The secure architecture for big data analytics is established by big data workgroup regulated by Cloud Security Alliance (CSA) [5] (shown in Fig. 5) (illustrated in Table 1). When used in healthcare scenario [21], EHR would become the input dataset to the general-purpose analytics platform having HDFS with FLUME and SQOOP at the data layer. All-time favorite MapReduce and HIVE perform the analysis on the data using machine learning algorithms by finding similar patterns. For storing multi-structured data HBase is used. STORM is a live streaming tool being used for any emergency coming up and intimating all the concerned healthcare staff immediately through AWS Lambda function. Intellicius and hunk produce the reports.

Data mining techniques used in healthcare industry benefits in terms of:

- Data modeling [21] with security [14] for healthcare systems
- Formulating treatment cost and resource availability as required

Table 1 Medical data sources as interpreted by various analytical tools (an illustration of Fig. 5)

Cite/tool	HDFS	MapReduce	Hive	Pig	STORM	Intellicius	Hunk
[21]	EHR is input to	For analyzing patients data	Analysis	Analysis	Live streaming	Reporting	Reporting

- Future behavioral prediction based on patient's history
- Devising of executive healthcare information system
- Public health informatics
- Implementing e-governance [5] structures in healthcare
- Health insurance [9].

Healthcare organization may implement Knowledge Discovery in Databases (KDD) [13] using frameworks defined in [20,23] with help of skilled resource who is well acquainted with medical jargon collected through different sources [6]. KDD [13] is a well-organized effort to determine previously unknown information from extracting useful data. KDD would be effective to determine a meaningful pattern in the data to develop strategic solutions [10] through effective data mining and machine learning steps [13] where specific algorithms are applied to extract patterns from data to further interpret.

Machine Learning and Data Mining overlap at some points during classification [13]. Data mining is perceiving the previously unknown information where its older sibling that is machine learning is known since the 1960s and focuses on classification and prediction. Machine learning algorithms evaluate or predict results based on previously known information utilizing it as a training set.

There are lots of critical questions regarding clinical assessment of patients that can be answered based on the relationship found by analyzing millions of patient's records stored in KDD. Questions can be such as:

- Based on the symptoms of patient what possible tests should be recommended for diagnosis?
- If these health-related problems occur, then which possible disease is most possible?
- What are the most common symptoms found in diabetic patients?
- What is the age group of Heart disease patients?
- Standard billing for a particular diagnosis is how much?

Data mining accomplishes tasks of class description, classification, association, prediction, clustering, and time series analysis. Online Analytical Processing (OLAP) [55] is one way to process data in a multi-dimensional capacity. Data mining techniques are illustrated [10].

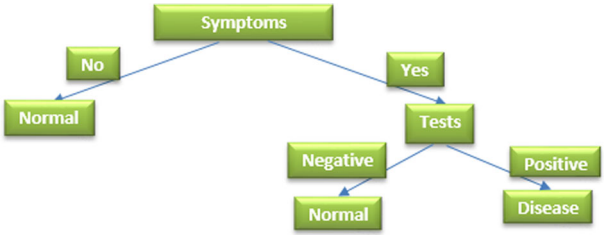

Fig. 6 Classification rules extracted from a decision tree for diagnosis

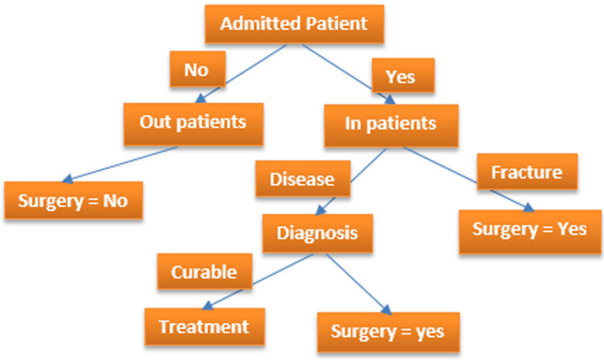

Fig. 7 Hunt's algorithm for inducing decision trees

4.1 Classification and clustering techniques

4.1.1 Rule-set classifiers

It extracts information by 'if-then' rules converting into knowledge. IF condition THEN conclusion.

IF part consists of one or more conditions using predictors and the rule consequent (THEN part) gives the result, in any case, depending on the predictor value. In healthcare (shown in Fig. 6), a common example is, if a certain symptom is found there is a list of some laboratory tests and in case, a certain test is found positive then a most appropriate disease is diagnosed after consultation of doctors.

$$\text{Symptoms} \longrightarrow \text{cause of disease}$$

4.1.2 Decision tree

Knowledge is represented by nodes and branches (Fig. 6). Every nonleaf node is labeled with attribute values in that node and the branches coming out of it assigns the values to these attributes. These trees as in Figs. 6 and 7 based models include association, clustering, classification and regression for common implementation of induction modeling. The most used decision tree algorithms [13] are C4.5, ID3, HUNTS, CART, SLIQ, SPRINT, etc.

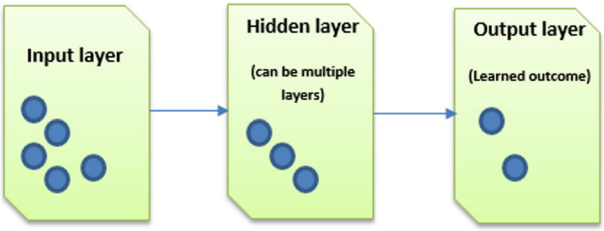

Fig. 8 A simple ANN diagram

4.1.3 Supervised/unsupervised learning

The learning is either supervised or unsupervised [3]. In supervised learning, the network is trained using a classification model based on given input and output pairs thus teaching it to relate such that it can predict output based on the input. Every time it matches the actual output with the desired and corrects the weights accordingly until there is no chance of further improvement and output is quite near or similar to the desired output. In unsupervised learning, the network learns only through inputs. It organizes itself in consistency with the group of similar stimuli such that an input space or a cluster is created having similar characteristics that match to the elements of real world. Hunt's algorithm is one that forms the basis for solving common longest subsequence problem (elaborated in Fig. 7 for healthcare services). In [20], an elaborate case study is given for modeling heterogeneous temporal EHR data using random subsequence based approach.

4.2 Neural networks algorithms

4.2.1 Artificial Neural Networks (ANN)

There is the concept of artificial neural networks (Fig. 8) in machine learning that work with computational paradigms based on mathematical models and mostly resemble a mammal brain, unlike any traditional computing. ANN is also termed as connectionist systems, adaptive systems, or parallel distributed systems [53]. They consist of adaptable interconnected elements called neurons with some weights assigned that simultaneously process and change with the flow of information and adaptive rules. ANN is primarily meant to mimic brain functionality in terms of computation as in cognitive processes through sensorial perception, categorization and association of concept and learning. ANN is an analytical process that uses cognitive learning for predicting the learned outcomes from previous observations. Neural networks in healthcare domain are applied in key areas like clinical diagnosis, signal analysis and interpretation, image analysis and interpretation, and drug development.

Today it is being used for applications to recognize patterns, classification, data compression and optimization. ANN behaves as a computational system having highly interconnected processing elements called neurons that are responsible for processing information as a response to external stimuli. An artificial neuron copies the behavior

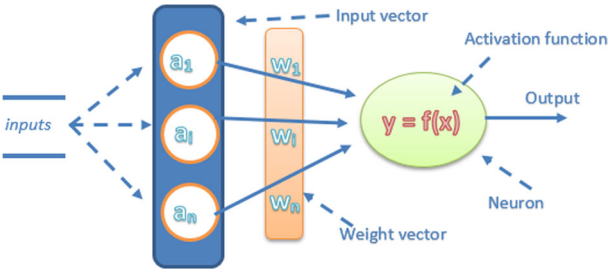

Fig. 9 Basic model of a single neuron [56]

of biological neurons through mathematical equations (Fig. 9). These neurons are bonded together through connections determining the flow of information in-between. The two types of signals [56] received by artificial neurons are (i) excitatory, and (ii) inhibitory to stimulate. If the input is excitatory the most likely signal that would be transmitted is to be excitatory as well. If the input signal is inhibitory most likely the propagation would be of inhibitory signal.

The ANN [56] exhibits in two modes: (i) learning or training, and (ii) testing. At first, when the training is at its initial stage, the network guesses the output for every example and with time as it learns it modifies such that it reaches a stable stage where it is able to give a satisfactory output. Learning [3] is adaptive where weights assigned to each neuron change as it observes to give a best possible output. Neural networks are known to solve problems as people in real-world scenario thus its application is also seen in healthcare domain for clinical diagnosis.

Papnet is a commercial application based on ANN to assist in screening of Pap (cervical) smears. A pap smear test is there to examine cells taken from uterine cervix to identify signs of precancerous and cancerous changes occurring. On early detection, there is 100% percent chance of cure for cervical cancer. Traditionally, it is a work of human eye to detect dozens of abnormal cells from 30,000 to 50,000 normal cells in pap smear through the microscopic lens in a laboratory. The best laboratories can miss up to 30% of abnormal cases. With Papnet [56], an improvement is seen in accuracy level for the screening of precancerous and cancerous cells in the cervix.

In a cardiovascular domain, neural networks were trained using ECG measurements of 1120 patients suffering from heart disease and 10,452 normal people without any history of heart attack. The resultant performance of neural network was compared to the widely used ECG interpretation program and diagnosis of an experienced cardiologist, and it was seen that neural networks were 15.5% more sensitive than the interpretation program and 10.5% more sensitive than the cardiologist for diagnosing abnormalities.

To predict breast cancer, there is a solution in form of Entropy Maximization Network (EMN). The prediction accuracy of this neural network was measured to prove that it was most accurate in predicting five-year survival of 25 cases that were studied using backpropagation learning algorithm.

It is seen that ANN are near to accurate in predicting various other chronic illnesses like coronary artery disease, myocardial infarction, brain disease (multiple sclerosis

and cerebrovascular disease), Down's Syndrome in unborn babies, benign and malignant ovarian tumors, liver cirrhosis or nephrotic syndrome, and inflammatory bowel disease.

The accuracy of these learning algorithms is determined by comparing to Cox regression (56%) and other classification algorithms. There are radial-based function networks, multi-layer feedforward network trained with backpropagation learning algorithm, Bayesian posterior probability distribution used in a neural network input selection, and adaptive resonance theory mapping neural network (ARTMAP) and logistic regression. Imaging [56] is an important technique for ANN pattern recognition to extract and identify important features from radiographs, MRIs, ECTs, etc. Neural networks are also a widely used technique for signal analysis [56] as to detect four ECG waveforms, Contingent Negative Variation (CNV), phosphorus (31P) magnetic resonance spectra (MRS), normal and abnormal heartbeats. Finally, National Cancer Institute (NCI) in the USA came up with a drug development solution through implementation of neural networks [56]. The network correctly classified 91.5% of anticancer agents (drugs) a per their usage and action.

4.2.2 Neuro-fuzzy

Stochastic backpropagation algorithm [11] is used for the construction of the fuzzy neural network. Firstly, we initialize the connections with random weights. Then we compute the input, output values and error rate. Thirdly, uncertainty is measured by calculating certainty measure (c) for each node to decide.

The trained network consists of three layers (Fig. 8): (i) input layer, (ii) hidden layer, and (iii) output layer. There are 5 input nodes with 3 hidden nodes and resulting 2 output node (Fig. 8). With appearing of thrombus or blood clot in 75% of the surface area of the lumen of an artery the cell death or heart disease is predicted as per medical guidelines [11].

4.2.3 Recurrent Neural Networks (RNN)

Doctor AI [12] is a model given to mimic the clinical behavior of physicians and develop a successful application of Recurrent Neural Networks (RNN) for jointly diagnosing the future disease with medical prescription with timings. The system takes all the medical history of patient to forecast and is tested on a real-world EHR data over 250k patients from past 8 years.

4.3 C4.5 algorithms and beyond

C4.5 produces classifiers in form of decision trees or also may be as rule-set form. There are multiple algorithms been used in C4.5 that are discussed here [13].

4.3.1 Decision trees

C4.5 [13] works as decision tree while using the divide-and-conquer algorithm. In case, there is S class having various attributes and thus, the most frequent attribute

makes the leaves of a tree. Another case is such where a test attribute (S) have several outcomes that are termed as S1, S2, S3, and so on that become nodes and this process goes on recursively.

Likewise, several such attributes make C4.5 differ from the CART. Tests in CART only give binary trees where C4.5 can give more than two outcomes. C4.5 also differs while indexing tests as it uses information based ranking criteria and CART uses GINI diversity method. CART prunes trees using cost-complexity method while C4.5 uses single-pass algorithm having binomial confidence limits. When a test value is unknown, the CART puts an estimated approximate value of the outcome but in case of C4.5, it puts apportions probabilistically.

4.3.2 Rule-set classifier

In Sect. 4.3.1, it is seen that C4.5 outperform other algorithms like CART as complexity is felt in constructing decision trees that are limited to binary trees alone and finding outcome or conclusion remains cumbersome. C4.5 [13] make use of several rules in form of (if A and B and C and … then class X). Rules for every class are grouped and a case is classified as getting the first rule with its conditions satisfied and if no rule gets satisfied the class is classified by default.

List of rules are formed by an unpruned decision tree. The path along the root to the leaf of a tree form the rules and its conditions are the outcomes along the path where leaf becomes the label for a class. A hill climbing algorithm drops conditions until the pessimistic error rate is found.

4.3.3 C4.5 learning algorithm

Till now a lot of work is done in coming up with many algorithms to classify and predict the chronic diseases like kidney disease [15], heart disease, breast cancer, motor neuron, and diabetes.

C4.5 data mining algorithm predicts kidney disease [15] using a learning algorithm as briefed in [3]. Using C4.5 learning algorithm classification is done based on two classes: (i) chronic kidney disease and (ii) not chronic kidney disease. The dataset was taken through UCI machine learning repository and platform were chosen was Weka thus comparing the classification model with ORANGE or TANAGRA. The hypothesis used were four: (i) True Positive (TP), positive samples correctly predicted, (ii) True Negative (TN), negative samples correctly predicted, (iii) False Negative (FN), positive samples wrongly predicted, and (iv) False Positive (FP), negative samples wrongly predicted as positive. Accuracy was measured of this classification model applied on 400 instances, it was 63% as out of 400 instances 396 were rightly classified and 4 were with an error. Resultantly, it was determined that C4.5 is an excellent learning algorithm [15].

4.3.4 See5/C5.0

C5.0 or See5 [13] superseded C4.5 in 1997. The new capabilities formed improved its efficiency in ways including (i) through boosting classifiers are voted to give final clas-

sification and improves predictive accuracy, (ii) new datatypes, variable misclassified costs, "nonapplicable" values, and mechanisms to prefilter attributes, (iii) in unordered rulesets, in case it is classified, all applicable rules are found and voted while improving interpretability of rulesets and predictive accuracy and (iv) C5.0 is able to take advantage of multiple CPUs to improve scalability through multi-threading.

4.4 Evolutionary algorithms

Evolutionary Computation [17,18,34] assist in providing several solutions to a field of medicine where problem is complex. This population of solutions evolve to ultimately reach to the most optimized solution. Most medical diagnosis is formed by searching a large and complex space like for cytologist analyzing a cytological specimen for deciding whether it is malignant or not, would have to search a complex space of all possible feature cells to separate a set of features for coming to the right diagnosis.

4.4.1 Synchronous evolutionary algorithms

Synchronism is achieved in an algorithm when all fittest nodes are sent and accepted at one time as in synchronous island model [17,19]. All islands wait till the last island undergoes its evaluation of subproblems to move on to next generation. It is thus known to slow down the throughput of an algorithm.

4.4.2 Asynchronous cellular evolutionary algorithms

To update the cells of a cellular Evolutionary Algorithms (cEA) [17] sequentially having a population in a 2-d grid there are many ways and the most general one is independent random ordering of updates in time while randomly choosing the next cell to be updated. Other three methods are fixed line sweep, fixed random sweep and new random sweep.

The asynchronous updating of the decentralized population in cellular Genetic Algorithm (GA) is tested by applying three alternative policies [17]. It is concluded that synchronous update policy is best in case of search grid where its success relates to the percentage of hits. However, in case of finding the optimum solution or reach maximum the efficiency the asynchronous methods [19] are faster. Line Sweep (LS) asynchronous policy is best for the convergence of two problems: MMDP and P-PEAKS with a high and desirable success rate to that of a synchronous update. Fixed Random Sweep is best suited to hard FMS problem. Fast convergence refers to local optima in evolutionary algorithms but cellular GAs in general.

5 Analysis of big data healthcare analytics platforms and algorithms

In Sect. 3, we have already done an in-depth study of various healthcare platforms that have been proposed. In Sect. 4, a very rigorous exploration is done over various algorithms being widely used in healthcare domain for prediction, diagnosis, and treatment using clustering and classification techniques using evolutionary algorithms

[18]. From the review of the literature, we know that mainly there are two approaches used for analysis of healthcare big data; these are data mining and machine learning techniques [54]. Under the umbrella of these two approaches there is a variety of algorithms in form of neural networks, fuzzy logic, synchronous and asynchronous evolutionary algorithms and finally, these algorithms can work in hybrid approach [16,43,44,49] to optimize.

Table 2 shows a detailed overview of various healthcare systems that have been proposed till now. It is seen that most of these systems are built over some widely used platforms known for big data analytics like Hadoop or Google big query platforms. The success parameter of these systems is seen to be compatibility with HIPAA [33] and HL7 standards. If not, then the highest accuracy is tried to be achieved. The measures for determining accuracy level also differ and may have some weaknesses. There are various other limitations been traced depending on the scope of the problem or the availability of datasets. There are many systems that have been visualized and the most prominent of them are those that dealt with big data, cloud or IoT based on some standard measures.

After a thorough analysis, the researchers have found in this study that the foundation and conceptualization of all these proposed healthcare systems lead to the future vision for learning healthcare system (LHS) presented in [3,4]. The research and the work over its implementation is being carried out in Mayo Clinic, Rochester, USA [34].

5.1 In-depth evaluation of healthcare big data analytics and frameworks

Careful review of several research studies about healthcare systems being evolved on top of some known platforms, algorithms, tools, and techniques is done to innovate further. Analyzing the big data spanning over hospitals, patient histories, physicians' prescriptions and diagnosis, diseases (general as well as chronic), and electronic medical records (EMR) is complex requiring lot of computation and intelligence.

Findings are evaluated through thorough exploration based on our extensive research study.

5.1.1 International standard (HL7) application on cloud-based healthcare platform

Predictive Analytical Decision Support System (PADSS) [22] integrates into a cloud-based healthcare platform that is Message Oriented Middleware (MOM). It connects healthcare organizations to share data using a customized Health Level Seven (HL7) platform having Fast Healthcare Interoperability Resources (FHIR) specification to provide a proper structural format to exchange data for prediction of disease outbreaks in real-time scenario. FHIR in HL7 enables clinical and administrative data exchange on international standards between healthcare applications. FHIR data model is extensible to allow applications to modify on need basis while using a set of customized data structures and resources.

Table 2 Analysis of various proposed Healthcare platforms

Cite	Proposed technique/system	Application	Algorithms/ platforms	Testing/validation technique	Accuracy level	Strength	Limitations
[31–33,35]	Healthcare Informatics and Big Data	Diagnosis, treatment, and follow-up of human disease	Big Data Taxonomy for Healthcare Informatics	New hypothesis in disease management is being formed for diagnosis, prevention and providing personalized treatments	Standardization	Work Group is formed to encourage researchers	Lacks actual transformation due to complexity
[7,30,36]	Health IoT Cloud Architecture	IoT can address pediatric and elderly care, chronic disease supervision, private health, and fitness management	IPv6, 6LoPAN protocols agents connected to data layer through application layer	Exploration into existing IoT-based healthcare technologies	State-of-the-art architectural design	Basic architecture is visualized	Not integrated with big data analytics
[8]	Health Twitter Big Data Management	Diagnosis and treatments of different diseases, habit changing and disease prevention, etc.	Hadoop framework using Mahout	The Twitter API with twitter4j library is integrated with web application using NoSQL database in MongoDB getting stream of twitter real health data as an input to form statistics and results	Results give useful insights for a common user as well as practitioners and doctors	Initiates decision support system for healthcare	Input Data stream limited to social media integration

Table 2 continued

Cite	Proposed technique/system	Application	Algorithms/ platforms	Testing/validation technique	Accuracy level	Strength	Limitations
[3,4,34, 46]	Learning healthcare system	Mayo Clinic has come up with a web-based application that is AskMayoExpert (AME)	Decision support system named MEA was implemented on top of big data empowered NLP infrastructure in its Unified Data Platform (UDP)	Three working Care Process Models (CPMs) that are hyperlipidemia, atrial fibrillation and congestive heart failure (CHF) in the pilot phase of MEA	Standardized	HL7 compliant	Not connected to IoT infrastructure
[25]	Synchronous Big Data Analytics	Telehealth services	PhysicalTherapy-as-a-Service (PTaaS) developed on top of Microsoft Kinect having motion sensors	Conducting various experiments and in a real-time scenario based in homes for elderly in Kansas City connected by Google fiber and clinic that was situated in Columbia, Missouri	High level of enthusiasm and acceptance is reported	Patients get remote therapy at home	Network configurations and time synchronization in online analytics was found challenging

Table 2 continued

Cite	Proposed technique/system	Application	Algorithms/ platforms	Testing/validation technique	Accuracy level	Strength	Limitations
[24,26]	GEMINI: an integrative Healthcare Analytics System-Later integrated with: Cloud-Based Patient Profile Analytics System	Patient profiling and analytics to predict unplanned readmissions, and chances of getting diabetes by observing family history and dietary plan	EPIC is there for managing scalability of self-learning knowledge base that keeps the healthcare specialists in loop for feedback to be updated using NLP processing and data analytics. Techniques used are feature selection, clustering, classification and prediction	A case study is presented predicting the risk of unplanned patients' readmissions. A tool used is Weka using Bayesian Network Classifier to run 10-fold cross-validation. Implemented	Precision = 0.388 Recall = 0.457 It is better than manually handled patients Improves with maturity	Self-learning knowledge base using feedback loop	Precision and recall should be better. challenge lies in moving big data in and out of the cloud for health institutions and organizations that is in terabytes and petabytes at the speed that is needed
[27]	Predictive Analytics System for Diabetic Mellitus (DM)	Predicting Diabetes Mellitus	Hadoop using MapReduce is the chosen platform	Conceptually explained	Efficient point of care	Designed more for rural areas to early predict the occurrence of DM	Experimental results are lacking
[22]	Cloud-Based Healthcare Platform	Decision Support System	Platform integrates Predictive Analytical Decision Support System (PADSS) built on top of PREVENT	Two simulations were deployed on Google Cloud Platform to access its responsiveness, scalability, and reliability	In compliance with HL7 with FHIR	Perform real-time large big data analysis jointly with Complex Event Processing (CEP)	Tuning up of SQL queries is required

Table 2 continued

Cite	Proposed technique/system	Application	Algorithms/ platforms	Testing/validation technique	Accuracy level	Strength	Limitations
[9]	Applying Big Data Analytics to Social Insurance Huge Information of Patients	Patients Insurance System	Hadoop framework	Scenario-based	Detect missing details of the patient	Is adaptive to be used for other novel data mining techniques for analyzing healthcare big data to investigate, restorative examination and clinical decision support for a specific patient	Limited in scope
[23]	Multi-Level Data Analysis (MLDA) Framework	Mobile diagnostic app	Neural network computations and multiple level clustering analysis, ubiquitous patient classification	RapidMinor is used for experimental setup of MLDA framework using DBSCAN, k-means and k-medoid algorithms	Multiple level DBSCAN was found better with overall similarity = 0.86	It gives good results for classifying heterogeneous big data	It is a standalone framework and is not standardized as well as not connected to IoTCloud infrastructure

Table 2 continued

Cite	Proposed technique/system	Application	Algorithms/ platforms	Testing/validation technique	Accuracy level	Strength	Limitations
[16]	Hybrid Intelligent System	Applied to diverse clinical scenarios, like; breast cancer diagnosis, diagnosis of coronary artery stenosis, analysis of microcalcification on digital mammograms, control of depth of anesthesia, and assessment of myocardial viability	(i) Low-level relay hybrid, (ii) low-level teamwork hybrid, (iii) high-level relay hybrid, and (iv) high-level teamwork hybrid	Taxonomy is presented	Sequential as well as parallel hybrid algorithms are compared and trend is seen toward parallel algorithms	It gives multiple hybrid approaches using relays as well as teamwork of agents based on scenario	It may not be covering all relationships of many hybrid algorithms

5.1.2 Healthcare informatics integrated with social networks

An opportunity [36] is realized in integrating social networks like Twitter and Facebook, etc., with Healthcare Informatics. mHealth technologies and social media connects patients outside the clinical premises to themselves as well as doctors. Therefore, gaining insights on various cures and precautions becomes cost-effective and easy.

Hadoop with Mahout is seen as a possibility to perform analytics on social data coming through Twitter and Facebook [8].

5.1.3 HealthCare analytics being employed for particular diseases

Complete architecture of Internet of Things (IoT) [7] connect wireless sensor devices attached to the body to detect health and diagnose early symptoms of any disease particularly chronic disease such as Heart Failure, and Diabetics while continuous monitoring for BP control and body temperature. There is risk arising in hypertension symptoms from these chronic diseases and detection is thus recorded from a chronic disease management program [52], i.e., Vanderbilt MyHealthTeam (MHT). Patients with lesser control over BP are found to have a risk of hypertension. Hypertension that is prevalent among over a billion patients all over the world and only in 65 million Americans forms the most common risk factor for cardiovascular disease. Further, as more diversified clinical data is analyzed the prediction of early onset heart failure (HF) is made possible [53] and these clinical applications are tried over 10k patients spanning over seven years. Another risk prediction model [54] assesses EHR to score risk of diabetic complications and Heart Failure making features out of patients' profiles.

Evolving learning healthcare system [3,34] is yet another possibility being considered where all healthcare stakeholders and diseases are linked. Applying inductive reasoning and other healthcare analytics researchers are moving toward precision medicine through decision support system for diagnostics and prognosis of various diseases.

5.1.4 Healthcare analytics systems employed specific to diabetes

It is understood that diabetes mellitus is an uncommon chronic disease. There are various healthcare analytics [27] for predicting diabetes and advising on its possible treatments for underlying complications.

In MLDA framework [23] as well the diabetes patients' data has been taken to perform various big data analytics approaches like Multiple Level DBSCAN and K-means through clustering, association rule mining and classification.

5.1.5 Types of datasets that form the input

There are various types of datasets (Table 3) in healthcare that are studied. Most common globally used dataset being built today is Electronic Health Record (EHR) having a standard coding system for all healthcare related jargon such as one is International

Table 3 Relating various models/systems and diseases employing some known healthcare datasets

Model/cite	[7,12,15,56–59]	[22,24,26,27]	[60]	[8,23]
Predictive	Diabetes, coronary artery disease myocardial infarction, Hypertension, brain disease (multiple sclerosis and cerebrovascular disease), heart failure, Down's Syndrome in unborn babies benign and malignant ovarian tumors, kidney disease, liver cirrhosis or nephrotic syndrome, BP, body temperature using ECG Measurements, national/electronic health record, clinical data and case histories	Predictive model is developed using real EHR datasets with up to 300,000 patients		Diabetes and cardiovascular patients data to predict disease and maximum work load they can take
Entropy Maximization Network (EMN)	Predict survival from breast cancer			
C4.5	Kidney disease			
Papnet	Pap smears			
PARAMO		Performance is assessed using real EHR datasets with up to 300,000 patients		
Kullback–Leibler divergence retrieval model			Medical literature retrieval model for case queries, literature articles, ImageCLEF based on physicians feedback	
Doctor AI	EHR, PAMF			
IoThNet	BP, body temperature, diabetes and heart failure using EHR and RFID			

Table 3 continued

Model/cite	[7,12,15,56–59]	[22,24,26,27]	[60]	[8,23]
Healthcare Social Network Analytics				Twitter/Facebook Data
Cloud-based Healthcare System		Disease surveillance using streaming data		
GEMINI		Monitoring Diabetes mellitus using EHR from patient profile graph		

Classification of Diseases revision 9 (ICD-9) [53] where codes for diagnosis, procedures and medications are listed [12]. EHR is a repository keeping all data not just related to diseases but also patient histories, diagnosis report, physician's notes, X-ray reports, data captured from RFID, medical and surgical instruments being used, medicines and pharmaceutical data. There are medicine-related literature articles that are kept in databases like PubMed [56]. The popular sources of patients' data to form training sets include Sutter Palo Alto Medical Foundation (PAMF) [12]. The whole Internet is connected via the Internet of Things (IoT) to form IoT HealthCare Network (IoThNet) gathering health data [7] from various devices and sensors through Radio-frequency Identification (RFID) to high-end PCs. Also, currently the data being analyzed has become vast and is collected from diverse geographical locations and sources [21].

5.1.6 Various detections and predictions have been perceived

It has been established that healthcare industry is producing vast amount of data that is now being termed as big data. This big data is being analyzed to make identification, detections and predictions (Table 4) studying patients' history, medications being given for various diseases (Table 4) and the results achieved [12]. Before giving various medicines for various chronic diseases doctors would know the side effects that could be suffered by the patient and take many precautionary measures [21]. Before diagnosing physicians detect [21] and study various symptoms carefully and conduct various tests to be sure of the diagnosis [12].

If required patients are called for readmissions [12,21] based on their case histories. Studying the resulting pattern of symptoms, diagnosis, the amount of drug dose and side effects formed can help further predictions and prevention [21] of diseases (Table 4).

Analysis of clinical data also helps to find out the risk factors that form the basis for major problem transitions [57] depending on the results gathered from de-identified patient cohort.

It helps identifying the whole operational level [21] of hospitals and treatment quality that is been maintained through the years. Proper expiry dates of surgical instruments are known and monitored.

5.2 Various models employing a set of proven algorithms, methods and classifiers using various tools

There have been a lot of work done in proposing various healthcare analytics models having a mix of algorithms following different methods for selecting various features from the data and predicting for future revolution in the field of medicine. As Scalable Orthogonal Regression (SOR) is an optimized nonredundant feature selection method is designed [58] having SVM classifier (Table 6) with Gaussian Kernel tool.

A survey conducted by Archenaa and Anita [21] recently in 2015 on big data analytics in healthcare reflects some known tools (Fig. 5) (Table 1) that have been used in such applications that are HDFS that takes input and further passes data to

Table 4 Identifications based on data input

Identification/Cite	[3,12,54]	[21,27,45]	[7,26,34,56,57,59,60]
Symptoms	To reach any diagnosis symptoms, patient's medical history and related disease form the best tools	Detection through case histories or performing predictive analytics to EHR	Rules to diagnose based on symptoms may be found in EHR, physicians feedback and medical thesauri
Diagnosis	Predicted using EHR	Medicine prescription with amount of drug dose based on clinical data	Of chronic diseases like cervical cancer, breast cancer, hypertension, cardiovascular and heart failure using national/electronic health record, clinical data, case histories, UCI data mining and machine learning repository
Prescribed/developed medication	Medication is prescribed based on EHR	PARAMO is a predictive model that may be used as clinical decision support system and medication may be prescribed based on similarity of patients' case histories	For drug development NCI used the pharmaceutical data to classify anticancer agents with 91.5% accuracy
Side effects	Diabetes has its side effects that need to be managed	Using clinical data	Hypertension is a side effect that is controlled
Readmissions	Predicted using EHR	If required using clinical data	Based on patient medical profile history future readmissions may be predicted with an estimated time interval
Prevention	Not much work done but tools are there to predict before time and take precautionary measures	Preclinical assessment would help determine risk in time by using social/geo data or EHR	Predictive analytics like ANN is good at solving complex problems and may be used for prevention of disease
Hospital quality	Health Informatics is composed of standardized clinical decision support system under governing bodies	Detection using national/electronic health record (EHR)	LHS administers quality governance
Treatment quality	Standardization is to ensure quality	Clinical data determines effective medicines	LHS administers quality governance
Problem transition	Side effects occur in for of problem transition	PARAMO is a full cycle model to monitor patients' EHR for problem transition	Hypertension is form of problem transition
Expiry dates of instruments	Administrative governance uses EHR to filter outdated/used instruments	Of medical and surgical instruments through EHR and RFID	LHS administers quality governance

MapReduce, pig or hive for analysis. STORM is used for live streaming and further Intellicius and Hunk do the reporting.

Major models (Table 5) that have been developed so far are Patient Risk Prediction Model, Kullback–Leibler Divergence Retrieval Model, Doctor AI, and IoT healthcare network (IoThNet). These models employ some algorithms in combination [57,59] having various methods (Table 6) for analyzing the vast data.

5.2.1 Patient risk prediction model

Patient Risk Prediction Model [54] is built using predictive model (e.g., classifier) and SVM that employs Top-K Stability Selection algorithm shown in Fig. 8 [54] and sparse logistic regression algorithm extended version of stability selection and sparse learning method with Lasso using feature selection and classification methods.

Algorithm: Top-k stability selection [54]
input: dataset T = {xi, yi}ni=1, iteration number β,
parameter set A
output: top-k stability scores SSS(k)(f) for all f ϵ F
1: for j=1 to β do
2: subsample D(j) from T without replacement
3: for λ ϵ Λ do
4: compute the sparse logistic regression on D(j)
using parameter λ and obtain the result ŵ(j)
5: store indices of selected features:
$\hat{S}\lambda(D(j)) = \{f: \hat{w}(j)f \neq 0\}$
6: end for
7: end for
8: for f ϵ F do
9: compute selection probability for all λ ϵ Λ:
$\Pi f\lambda = 1/\beta\ \Sigma\beta i=1\ 1\ \{f \epsilon\ \hat{S}\lambda(D(i))\}$
10: compute top-k stability score
$SSS(k)(f) = \Sigma\lambda$: $\Pi\lambda f$ ranks top-k$\Pi\lambda f/k$
11: end for

5.2.2 Kullback–Leibler divergence retrieval model

Kullback–Leibler Divergence Retrieval Model [14,56] with Dirichlet Smoothing is a state-of-the-art method used for retrieval. Using pseudo-relevance method that treats the top-ranked documents as the relevant documents extracting useful terms from these feedback documents and further the rare keywords are matched against query rather more frequent terms but it is not sufficient so the related or semantically distinguished keywords relevant to medicine or disease are given higher weights that influence the MeSH and UMLS dictionaries for further improvement and upgrade. The test set

Table 5 Relating various models with their respective algorithms

Model/algorithm	[12]	[7]	[56]	[57]	[59]	[60]
Prediction model	Recurrent Neural Network (RNN)			Random Forest, logistic regression, naive Bayes	Lasso, sparse logistic regression, top-k stability selection	
C4.5						
Kullback–Leibler divergence retrieval model						Dirichlet smoothing
Doctor AI	Recurrent Neural Network (RNN)					
IoThNet		Wireless sensor networks				
Papnet				Artificial Neural Networks (ANN)		
Entropy Maximization Network (EMN)				Artificial Neural Networks (ANN)		

Table 6 Algorithms employed with respect to their methodology

Cite/model	LASSO	Feature selection	SOR	The high precision retrieval method	The sparse learning method	Top-k stability selection	Gaussian kernel
[58]	Lars	Greedy, filter, wrapper approach	Predictive, Feature Selection, LASSO	Information Gain, $\chi 2$, mRMR, area under ROC curve			SVM
[59]		Sparse Logistic Regression used by			Sparse Logistic Regression	SVM, Random Forest, Logistic Regression	

prepared from ImageCLEF medical case retrieval dataset is evaluated applying high precision retrieval method over it within the framework of Kullback–Leibler Divergence Retrieval Model. Expected performance precision that is set for the proposed model is 0.3980 but experiments showed actual results having precision around 0.2754 that would improve over 40%.

5.2.3 Doctor AI

Over the time a large amount of EHR data is collected having records of patients treated by various physicians in terms of procedures, diagnosis, and medical prescriptions. Understanding this data as modeled in [27] a general pattern of how physicians behave in various scenarios is visualized and the model that is formed is known as Doctor AI [12]. It incorporates a successful application of Recurrent Neural Networks (RNN) employing Theano tool to forecast future diagnosis and medical prescription artificially at their time. Doctor AI registers multiple clinical events and learns the dynamics of patient's diagnosis (Dx), prescribed medication (Rx), and the revisit that is expected. The model is fed with ICD-9 codes to correctly articulate results. Classifier used with RNN is GRU. And for conducting experiments the test set is created using the patient data from Sutter Palo Alto Medical foundation (PAMF) limited to heart failure cases over the span of 8 years.

5.2.4 IoT healthcare network (IoThNet)

Now we have entered into the era of IoT where the electronic devices and commodities are connected Internet or cloud through sensors, RFIDs, wireless devices, etc., known as smart objects. Being prevalent in various application industries it specifically focuses on healthcare domain and is a hot area for research and development. IoThNet on cloud has its own topology, architecture and platform. When all the healthcare-related wireless devices and sensors connect to cloud make a topology. The patients all over the region connect through devices and communicate over the framework to data layers kept in cloud servers. The whole IoThNet architecture is built over multiple layers in a wireless sensor network (WSN) using various protocols as a medium among which IPv6 and 6LoWPAN form the basis of IoThNet [7]. The platform combines the network platform and computing platform together over various healthcare business layers shown in Fig. 10.

5.2.5 Hybrid intelligent system

Here, the hybrid framework [16] is referenced that is adaptable and evolve for future exploitation to give us future Healthcare analytics investigative system, i.e., SmartHealth [21], for learning healthcare system.

6 Limitations and strengths

The analysis of big data itself becomes the bottleneck because of its massiveness. Therefore, the healthcare domain experts are looking toward computer specialists to

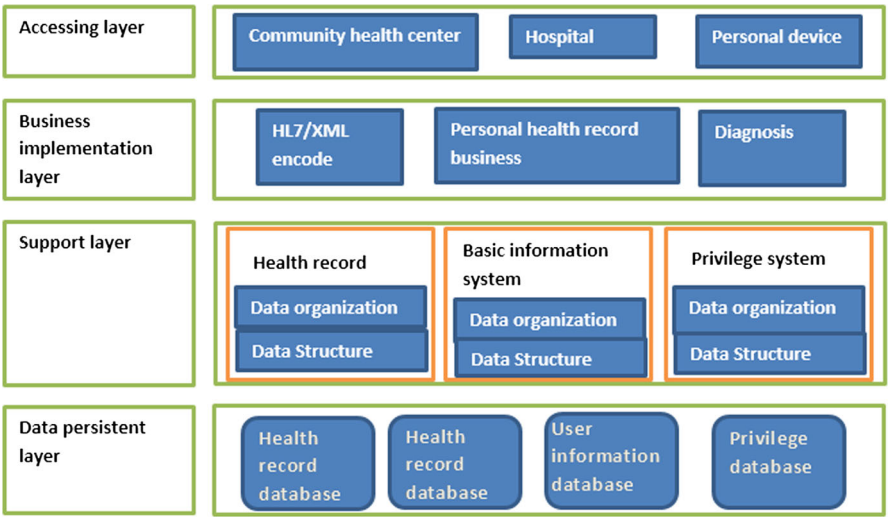

Fig. 10 Functional framework of a health information service model [7]

investigate and come up with the solutions to learn through data. There are some emergent technologies in the field of data science: Hadoop, unsupervised learning (finding hidden patterns in data), graph analytics, and natural language processing (where knowledge is extracted from documents enabling computers to understand the textual language of humans). The big heterogeneous data in the medical field has overwhelmed the intuitive abilities of the research community. The need of algorithms is highly recognized through which correlations are developed been all the associated factors and features. Thus, a massive opportunity lies ahead promoting medicine as an information science laying the foundation for learning healthcare system [3].

There are many challenges needed to be addressed [35]. The massiveness of big data in healthcare itself is a big issue. Looking into the US healthcare system only five years ago gives us an estimation of data reaching 150 Exabyte (1018). It would not take long that we would have to deal with data amounting to zettabytes (1021) and yottabytes (1024) when other countries also get into the picture. Data are coming in from many platforms that are mainly social media, real-time imaging, high throughput sequencing platforms, the point of care devices, mobile health technologies and wearable computing. Big data has six Vs that also apply to healthcare data: value, volume, velocity, variety, veracity, and variability. Big data is itself a challenge the other challenges occur due to heterogeneity and variety in data that is coming from various mediums.

The systemized and accurate view of information about patients to clinicians is not apparent and the knowledge of risks and benefits is often vague. Even if the evidence exists for a particular decision in a specific case it is not always applicable to the patient [3].

There is not much difference in the algorithms and models used in traditional healthcare systems and that of big data. The great difference is felt in traditional

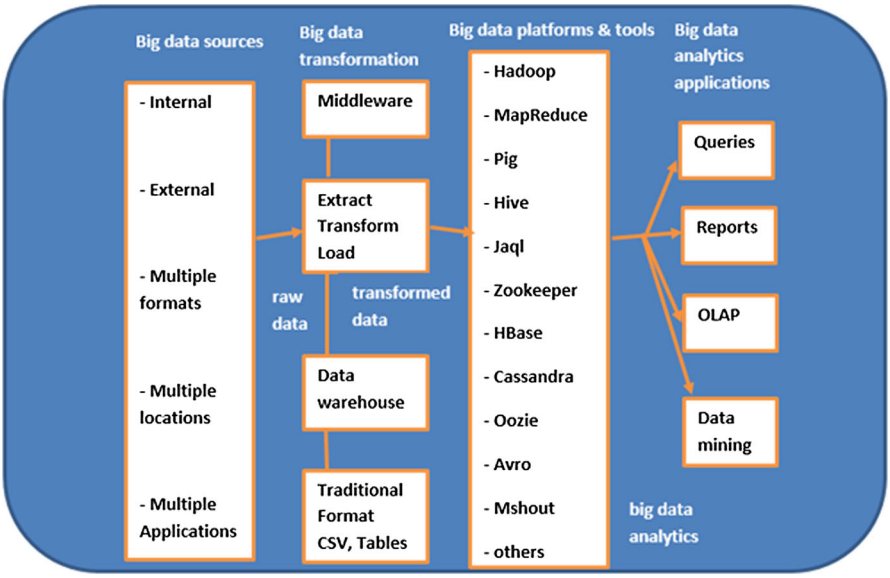

Fig. 11 Conceptual architecture of big data analytics [1]

analytics tool and those used for big data analytics with respect to user interface. Big data analytics tools are not as user-friendly as traditional tools have become. There is extensive programming involved, are complex, and requires a variety of skills. Big data analytics tools are open source and thus are not as sophisticated as proprietary tools. In big data analytics, complexity begins with data itself until it is processed and analyzed (Fig. 11).

There is a spread of vast unstructured data coming from multiple nodes that can be a great distance apart and need not be authentic. It is a very challenging job to map this unstructured data on to HL7 coding standard [33] as there is a high need for an analytical tool that serves solely to the healthcare industry. Standardization [7] is a necessity as many researchers and vendors are promoting several eHealth services with biotechnology but there is still need for collective recognition from standardization bodies like Information Technology and Innovation Foundation (ITEF) and Internet Protocol for Smart Objects (IPSO). Also, there are several security issues [7] still prevalent in the system and call for challenging research breakthrough.

Even with the advent of IoThNet [7], the concern remains of any malicious user intervention or a misleading health diagnosis because of various factors that may get manipulated in the process resulting to risk one life.

Healthcare analytics are more sensitive to their nature as they are concerned with lives of human beings. The results from PTaaS [25] clearly demonstrated the challenges for network configurations and time synchronization in online analytics applied on PTaaS data streams. This is a basic limitation of using synchronous big data analytics. So, the researchers have to be very cautious in their presumptions and resulting accuracy (Tables 1, 7) of these tools and algorithms.

Table 7 Accuracy comparisons for various algorithms

Algorithm/accuracy [12,57,58,60,66]	91.30%	AUC (Area Under Curve)	0.398 average precision	79% recall@30
ARK	Achieved by [66]			
SOR		Proportional (as more features are selected AUC increases) [58]		
Transition point prediction		0.7 approx. [57]		
Kullback–Leibler divergence retrieval model			Achieved by [60]	
Doctor AI				Achieved by [12]

It is not at all easy to comply with the needs of stakeholder in healthcare whether they are patients, medical staff, administration and other governing bodies. And, it is already late for the adoption of big data approaches that has left the medical field not fully prepared for its aspiration of future precision medicine [3] to be designed for individualized healthcare using personalized information through learning from physicians and practitioners feedback [60].

For researchers [1], there is challenging opportunity lying in healthcare data analytics to study this vast amount of data, find patterns and trends within data and provide solution for improving healthcare, save lives while lowering cost. Big data analytics applications in healthcare have advantage of this explosive data to extract insights and support informed decisions. In healthcare domain, previous study [23] shows that there are many correlated diseases affecting groups of patients. Like, diabetic patients have chances of forming underlying diseases for liver, heart, kidney, etc. To understand and analyze these complex correlations in heterogeneous data is thus challenging.

Our motivation to make learning healthcare system (LHS) [4] a reality becomes our strength. It is the era when the use of open source platforms has become highly recommended and Hadoop/MapReduce are available on the cloud even as healthcare data analytics. There has been some work done as for streaming the healthcare-related data over the cloud realizing the need. Considerable solutions in form of IoT devices and healthcare social networks are developed. Applications like WebMD [63] was a start toward building more computationally intelligent systems. Global standardized conventions are formed that can comprehend data to the analytics tool being developed in its right form. Thus, there has been considerable effort being put in to systemize the healthcare system to be fed into the healthcare analytical tool. The architecture of IoThNet [7] is greatly studied and its efficiency is realized with effective integration of biotechnology and smart objects with the Health IoT Cloud.

Another observation is that Doctor AI [12] achieves up to 75% recall@30, higher than many baselines (Fig. 11). Different algorithms included in RNN based Doctor AI give results as shown in Fig. 11. Results are accumulated in three scenarios (i) when only predicting disease codes (Dx), (ii) only medication codes (Rx), and (iii) jointly predicting Dx and Rx with time for next visit. The results in Table 2 in [12] prove the proposed system is stronger than several baseline algorithms.

While retaining the complexity of patients and medical decision making big data approaches highlights the interactions between all the factors underlying to reach future of precision medicine diminishing the effects of associated risks and outcomes [3]. The design of cloud-based PADSS [22] in compliance to HIPAA standards confirmed the improvement in rate of prediction and analysis of disease occurrence for more precise decision making. In MLDA [23] framework a two-tier architecture approach is used where ubiquitous patient classification is achieved through mobile app labeling unlabeled examination histories and using multi-level cluster analysis. This knowledge is efficiently and effectively made available to various user profiles whether they are patients or medical staff. We are also observing social networks [8] becoming part of Healthcare Informatics. Researchers are good at finding patterns without knowledge of the outcome [3].

7 Accuracy model

Statistics in medicine [64] is being applied for estimating the accuracy of diagnostics in the context of multi-category classification. The definitions of net reclassification improvement (NRI) and integrated discrimination improvement (IDI) are extended further. Numeric characterizations for improvements in accuracy for binary diagnostic tests were done to prove betterment over analysis performed on receiver operating characteristics (ROC) curves or other regression approaches. Estimation and inference rules for multi-class NRI and IDI are generated along with two medical examples with asymptotic distribution results and alternative accuracy model is proposed to hyper-volume under the multi-dimensional ROC manifold (HUM) based analysis. Diagnostics and classification tasks are carried out in medical practice to rightly distinguish the status of patients [64] as disease-present or disease-absent. Classification sometimes involve more than two categories as in the example taken classification for synovitis had to deal with patients from five different disease categories requiring different patient management strategy, respectively. Biomarkers are used to predict the status of patient disease to follow other procedures accordingly. And statistical tools employed give insight over the level of diagnostic accuracy. Here, Table 1 in [64] shows the results of 1000 Monte Carlo simulations for giving S and R true values of NRI and IDI for computing diagnostic accuracy found in six cases.

With the data generated the estimation uncertainty was observed [50] for which there were alternative approaches like □2 test distributions under the null for two-category classification and the other could be bootstrap.

Table 2 in [64] shows a comparative analysis for employing HUMs with three tissue biomarkers of Synovitis X1, X2 and X3 where X2 is found to be a most accurate markup in terms of accuracy as it classified 65% samples accurately.

Firstly in [64], the diagnostic accuracy for the separate component was measured then these were tried in combination using multinomial logistic regression for further improvement in accuracy.

The other example shown in Figure 2 in [64] took Leukemia patients' data for classification divided into three categories of gene expression that are (i) acute lymphoblastic leukaemia (ALL) arising from T-cells (ALL T-cells), (ii) ALL arising from B-cells (ALL B-cells), and (iii) acute myeloid leukaemia (AML). The accuracy of biomarkers is evaluated [50] for rightly distinguishing the three classes of data.

In addition to testing multinomial logistic regression model, support vector machine (SVM) was also experimented [64]. Multinomial logistic regression was found better to interpret the coefficients and puts light on the efficiency of marker's response. It is observed that with the small sample size of NRI and IDI-based data, the efficiency of SVM decreases. Also, the evaluation of NRI and IDI would be based on correctly collaborated models when the old and new models are not nested.

8 Solution foreseen

There is a need for analyzing the risks involved with the implementation of already evolved algorithms, tools, and procedures, IoThNet [7] being the most recent evolution

that can serve in the implementation of LHS [4]. For the success of IoThNet, there has to be a solid and secure wireless network in place ensuring the data-level security. The first concern should be to make this data structured by employing a single platform that is jointly supported by the organizations like WHO [31] that is well versed with internationally recognized standard healthcare conventions being used. After that comes the turn for developing a standardized healthcare analytical tool that is linked to HL7 [30,33] coding standard to ensure the accurate resultant healthcare diagnosis, prescriptions, maintaining of histories and delivering the all kind of medical treatment on the doorstep of any person. It was recognized in PADSS [22] that tuning up of SQL queries would lead to better results in time constrained scenarios. When dealing with large heterogeneous data it is better to setup an upper limit for the number of delimited risk zone points in the map for filtration of results. Lastly, to achieve quality in results the profile settings need to be done for use of adaptive rules for effective decision making. In the MLDA framework [23], a two-tiered architecture approach is used where ubiquitous patient classification is achieved through mobile app labeling unlabeled examination histories and using multi-level cluster analysis this knowledge is efficiently and effectively made available to various user profiles whether they are patients or medical staff. Compliant to previous architectural methodologies there is need to make an underlying hybrid architecture [16] that is asynchronous [17] in nature and governed by international standards for healthcare analytics to form the basis of cloud-based universal solution that would be validated using the statistical accuracy model [64] over biomarkers applied.

An opportunity is realized in integrating traditional health informatics with mobile health and social health [35,36] over IoThNet cloud platform [7] enabled with analytics [55,65]. Furthermore, it [29] would be built in perspective to become part of LHS [4,34]. Mobile and social health platforms are there to connect patients with doctors outside the clinical premises. In social health, the communication is expanded from between patient and doctor to patient to patient as well. Patients with acute and chronic diseases like diabetes, cancer, and heart disease are communicating over the social networks to discuss their experiences with each other. There are public health surveillance systems as well in place that correlate the possible emergence of asthma-related attacks in a polluted environment. Mobile messaging is also being considered as a preventive therapeutic treatment for patient behavioral changes or in case of diabetes enforcing a possible change in lifestyle for positive results.

9 Conclusion and future work

In this survey paper, we were focused primarily on innovative big data analytics platforms already been developed and tested on some healthcare systems with solutions for patients suffering from various diseases. The critical issues that are associated with the large user-generated data are known. Now, as we have greatly studied various data mining and machine learning techniques that are used to form various healthcare analytics systems having their advantages with some limitations and shortfall it is concluded that we would focus on one disease in future. We would narrow the scope taking data of diabetic patients and the underlying diseases they are prone to like liver cir-

rhosis. In this way, a more focused approach would help us to mitigate the limitations that have been felt in previous work. Our approach would be to develop a context aware cloud-based platform SmartHealth [29] integrated with IoThNet embedding healthcare analytics for diagnosing diabetic patients with inherent underlying disease of liver cirrhosis. Our system is focused to achieve high intelligence while inheriting the positive aspects of the previous systems that have been studied in this paper.. To achieve high level of user acceptability we would make this framework in compliance with the latest Health Insurance Portability Act (HIPAA) standards with support of organizations like World Health Organization (WHO) [31]. Our system [29] would be targeted toward becoming an integral part of LHS [4] in future.

References

1. Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2(1):3
2. Perer A (2012) Healthcare analytics for clinical and non-clinical settings. Proceedings of CHI Conference
3. Krumholz HM (2014) Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood) 33(7):1163–1170
4. The Learning Healthcare Project the Learning Health Care Project. http://www.learninghealthcareproject.org/section/background/learning-healthcare-system
5. Cloud Security Alliance (2013) Big data analytics for security intelligence. Big Data Working Group
6. IBM Centre for applied insights (2014) Raising the game: The IBM business tech trends study
7. Islam SR, Kwak D, Kabir MH, Hossain M, Kwak K-S (2015) The internet of things for health care: a comprehensive survey. IEEE Access 3:678–708
8. Cunha J, Silva C, Antunes M (2015) Health twitter big bata management with hadoop framework. Procedia Comput Sci 64:425–431
9. Basuthkar VS, Srinivas C (2016) Cost effective knowledge based quality and value data extraction from clinical healthcare data. Int J Adv Res Comput Commun Eng 5(4):1098–1103 [India]
10. Kaur H, Wasan SK (2006) Empirical study on applications of data mining techniques in healthcare. J Comput Sci 2(2):194–200
11. Srinivas K, Rani BK, Govrdhan A (2010) Applications of data mining techniques in healthcare and prediction of heart attacks. Int J Comput Sci Eng 2(02):250–255
12. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J (2016) Doctor ai: predicting clinical events via recurrent neural networks. In: Machine Learning for Healthcare Conference, pp 301–318
13. Wu X et al (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37
14. Buczak AL, Guven E (2016) A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutor 18(2):1153–1176
15. Boukenze B, Mousannif H, Haqiq A (2016) Predictive analytics in healthcare system using data mining techniques. Comput Sci Inf Technol 1:1–9
16. Talbi E-G (2002) A taxonomy of hybrid metaheuristics. J Heuristics 8(5):541–564
17. Alba E, Giacobini M, Tomassini M, Romero S (2002) Comparing synchronous and asynchronous cellular genetic algorithms. In: International Conference on Parallel Problem Solving from Nature, pp 601–610
18. Ramesh AN, Kambhampati C, Monson JR, Drew PJ (2004) Artificial intelligence in medicine. Ann R Coll Surg Engl 86(5):334
19. Bujok P (2013) Synchronous and asynchronous migration in adaptive differential evolution algorithms. Neural Netw World 23(1):17
20. Zhao J, Papapetrou P, Asker L, Boström H (2017) Learning from heterogeneous temporal data in electronic health records. J Biomed Inform 65:105–119
21. Archenaa J, Anita EM (2015) A survey of big data analytics in healthcare and government. Procedia Comput Sci 50:408–413
22. Neto S, Ferraz FS (2016) Disease surveillance big data platform for large scale event processing. In: Proceedings on the International Conference on Internet Computing (ICOMP), p 89

23. Xiao X (2016) Data mining techniques for complex user-generated data. Politecnico di Torino, Turin
24. Ling ZJ et al (2014) Gemini: an integrative healthcare analytics system. Proc VLDB Endow 7(13):1766–1771
25. Calyam P et al (2016) Synchronous big data analytics for personalized and remote physical therapy. Pervasive Mob Comput 28:3–20
26. Kulkarni SM, Babu BS (2015) Cloud-based patient profile analytics system for monitoring diabetes mellitus. In: International Conference on Computational Systems for Health & Sustainability (CSFHS), IJITR, pp 228–231
27. Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun J (2014) PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. J Biomed Inform 48:160–170
28. Kobielus J, Marcus B (2014) Deploying big data analytics applications to the cloud. In: The Cloud Standards Customer Council 2014
29. Shafqat S, Abbasi A, Amjad T, Ahmad HF (2018) SmartHealth simulation representing a hybrid architecture over cloud integrated with IoT: a modular approach. In: Presented at the Future of Information and Communications Conference (FICC) 2018, Singapore
30. Health Level Seven Standard. American National Standard Institute. www.hl7.org
31. WHO Who we are, what we do. WHO. http://www.who.int/about/en/
32. Standards Organizations for the NHII. ASPE, 26-Nov-2016. https://aspe.hhs.gov/standards-organizations-nhii
33. ANSI Approved Standards. http://www.hl7.org/implement/standards/ansiapproved.cfm?ref=nav
34. Kaggal VC et al (2016) Toward a learning health-care system-knowledge delivery at the point of care empowered by big data and NLP. Biomed Inform Insights 8(Suppl 1):13
35. Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang G-Z (2015) Big data for health. IEEE J Biomed Health Inform 19(4):1193–1208
36. Cortés R, Bonnaire X, Marin O, Sens P (2015) Stream processing of healthcare sensor data: studying user traces to identify challenges from a big data perspective. Procedia Comput Sci 52:1004–1009
37. Escaravage J, Guerra P (2013) Enabling cloud analytics with data level security. In: Tapping the full potential of big data and cloud, Booz, Allen, Hamilton
38. Cao P et al (2015) Towards an unified security testbed and security analytics framework. In: ACM, Urbana, USA
39. Pentland A, Reid TG, Heibeck T (2013) Revolutionizing medicine and public health. Report of the Big Data and Health Working Group. World Innovation Summit for Health, Doha
40. Berkman LF (2001) Social integration, social networks, and health. In: Smelser NJ, Baltes PB (eds) International encyclopedia of the social & behavioral sciences. Pergamon, Oxford, pp 14327–14332
41. Valente TW (2010) Social networks and health: models, methods, and applications. Oxford University Press, New York
42. WHO Report from the first consultation of the health and social protection action research & knowledge sharing (SPARKS) network. WHO. http://www.who.int/tb/publications/sparksreport/en/. Based on: http://www.who.int/healthinfo/15_Social_Networks_Berkman_ok.pdf
43. Sengur A, Turkoglu I (2008) A hybrid method based on artificial immune system and fuzzy k-NN algorithm for diagnosis of heart valve diseases. Expert Syst Appl 35(3):1011–1020
44. Zheng B, Yoon SW, Lam SS (2014) Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst Appl 41(4):1476–1482
45. Khaing HW (2011) Data mining based fragmentation and prediction of medical data. In: Computer Research and Development (ICCRD), 2011 3rd International Conference on, vol 2, pp 480–485
46. Sawacha Z, Guarneri G, Avogaro A, Cobelli C (2010) A new classification of diabetic gait pattern based on cluster analysis of biomechanical data. SAGE Publications, Los Angeles
47. Phanich M, Pholkul P, Phimoltares S (2010) Food recommendation system using clustering analysis for diabetic patients. In: Information Science and Applications (ICISA), 2010 International Conference on, pp 1–8
48. Antonelli D, Baralis E, Bruno G, Cerquitelli T, Chiusano S, Mahoto N (2013) Analysis of diabetic patients through their examination history. Expert Syst Appl 40(11):4672–4678
49. Purwar A, Singh SK (2015) Hybrid prediction model with missing value imputation for medical data. Expert Syst Appl 42(13):5621–5631

50. Polat K, Güneş S, Arslan A (2008) A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine. Expert Syst Appl 34(1):482–487
51. Karan O, Bayraktar C, Gümüşkaya H, Karlık B (2012) Diagnosing diabetes using neural networks on small mobile devices. Expert Syst Appl 39(1):54–60
52. Menshawy ME, Benharref A, Serhani M (2015) An automatic mobile-health based approach for EEG epileptic seizures detection. Expert Syst Appl 42(20):7157–7174
53. Miah SJ, Hasan J, Gammack JG (2017) On-cloud healthcare clinic: an e-health consultancy approach for remote communities in a developing country. Telemat Inform 34(1):311–322
54. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I (2017) Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 15:104–116
55. Demirkan H, Delen D (2013) Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud. Decision Support Systems, vol 55, no 1, pp 412–421
56. Sordo M (2002) Introduction of neural networks in healthcare. Open Clinical: Knowledge Management for Medical Care
57. Sun J et al (2014) Predicting changes in hypertension control using electronic health records from a chronic disease management program. J Am Med Inform Assoc 21:337–344
58. Luo D, Wang F, Sun J, Markatou M, Hu J, Ebadollahi S (2012) Sor: scalable orthogonal regression for non-redundant feature selection and its healthcare applications. In: Proceedings of the 2012 SIAM International Conference on Data Mining, pp 576–587
59. Zhou J, Sun J, Liu Y, Hu J, Ye J (2013) Patient risk prediction model via top-k stability selection. In: Proceedings of the 2013 SIAM International Conference on Data Mining, pp 55–63
60. Sondhi P, Sun J, Zhai C, Sorrentino R, Kohn MS (2012) Leveraging medical thesauri and physician feedback for improving medical literature retrieval for case queries. J Am Med Inform Assoc 19(5):851–858
61. Eswari T, Sampath P, Lavanya S (2015) Predictive methodology for diabetic data analysis in big data. Procedia Comput Sci 50:203–208
62. Esfandiari N, Babavalian MR, Moghadam A-ME, Tabar VK (2014) Knowledge discovery in medicine: current issue and future trend. Expert Syst Appl 41(9):4434–4463
63. WebMD. www.webmd.com
64. Li J, Jiang B, Fine JP (2012) Multicategory reclassification statistics for assessing improvements in diagnostic accuracy. Biostatistics 14(2):382–394
65. Cloud Analytics Platform. Gurucul Predictive Security Analytics
66. Kang U, Tong H, Sun J (2012) Fast random walk graph kernel. In: Proceedings of the 2012 SIAM International Conference on Data Mining, pp 828–838