International Journal of Research in Engineering, Science and Management
Volume-2, Issue-2, February-2019
www.ijresm.com | ISSN (Online): 2581-5792

Survey of Data Driven Medical Treatment Suggestion Systems

M. Hariesh Ramanathan
UG Student, of Biomedical Engineering, Rajalakshmi Engineering College, Chennai, India

***Abstract*: The medical industry is experiencing an ever increasing number of affected individuals along with rapid and exponential growth of population. Using suggestion systems to assist the findings of a physician and automate process of analyses in healthcare indeed has a drastically wide scope, by providing accurate and trustworthy disease risk diagnosis prediction and medical treatment based suggestions along with data from electronic medical records, wearable technology and to monitor health and performance of individual apart from accelerated development in IoT. The healthcare sector is generating a vast amount of clinical data which is of great value when properly used. Properly analyzing clinical documents about patient's health anticipate the possibility of occurrence of various diseases. In addition, acquiring information from specialists of the regarding particular disease as per the requirement facilitates proper and efficient diagnosis. There are several existing systems available to diagnosis and recommend appropriate therapeutics. Some system uses various clustering based algorithms whereas some uses various medical test prognosis, automatic report analysis, and treatment recommendation irrespective of various stages of the disease. Apache Spark is a unified analytics engine for large-scale data processing which processes the results 100 times faster than Hadoop. The goals of high performance and low latency response can be implemented using the Apache Spark cloud platform.**

***Keywords*: AI in Medicine, Apache Spark, Big Data, Data Science, Hadoop, Medical Data Analytics, Machine Learning.**

1. Introduction

A medical data processing system is indeed a complex and very different from other work environments. Health care treatment advisories has now become a major industrial sector. India is currently experiencing an exponential growth in the medical sector. There are several existing systems available to diagnosis diseases and recommend appropriate treatment and therapeutics. But the accuracy and latency responses are highly inappropriate making them less applicable. To overcome this, a hybrid clustering algorithm is devised, which runs in the apache spark. Analytics provide an approach for decision making through statistics and research to discern patterns and quality of performance. Cloud analytics is a cloud-enabled solution that allows to perform business analytics. Cloud computing is the remote provision of computing services. Distributed computing is a concept that refers to multiple computer systems working on a single problem. Hadoop is an open source project that seeks to develop software for distributed computing. Hadoop distributed file system is a powerful distributed file system that provides high-throughput access to application data. Data driven medical treatment suggestion system, a proposed model uses 'Apache spark' a fast and general-purpose cluster computing system. Spark uses disk for processing, whereas Hadoop's MapReduce is disk-based. This results in faster and efficient processing.

2. A Brief insight to Big Data Analytics

A. Big data

The above section says how to prepare a subsection. Just copy and paste the subsection, whenever you need it. The numbers will be automatically changes when you add new subsection. Once you paste it, change the subsection heading as per your requirement. Big data is a term that describes a large volume of structured, semi-structured and unstructured data that has the potential to be mined for information. It also refers to data sets that are too large or complex for traditional data-processing application software to adequately deal with. In most scenarios the volume of data is too big or it moves too fast such a data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently. E.g. An example of big data might be petabytes (1,024 terabytes) or Exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources. The data is typically incomplete and inaccessible often it is loosely structured. It can be categorized into three forms as Structured, Unstructured and Semi-Structured.

1) Structured

Structured Data is used to refer to the data which is already stored in databases, in an ordered manner. It accounts for about 20% of the total existing data. Over the period of time, talent in computer science have achieved greater success in developing techniques for working with such kind of data and also deriving value out of it. However, issues arise when size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabyte.

2) Unstructured

While structured data is any data with unknown form or the structure, they have no clear format in storage. It accounts for about 80% of the total existing data, most of the data a person encounters belongs to this category. un-structured data poses multiple challenges in terms of its processing for deriving value

out of it. An example of unstructured data is, a heterogeneous data source containing a combination of simple text files, images, audio and video files, photos, GPS data, medical files, instrument measurements etc.

3) Semi structured

The line between unstructured data and semi-structured data has always been unclear, since most of the semi-structured data appear to be unstructured at a glance. Information that is not in the traditional format as structured data, but contain some organizational properties which make it easier to process, are included in semi-structured data. For example, NoSQL, XML documents are considered to be semi-structured, since they contain keywords that can be used to process the document easily.

B. Characteristics of Big data

Big data is often characterized by the 3Vs: the extreme volume of data, the wide variety of data types and the velocity at which the data must be processed. Doug Laney published a report in 2001 describing the characteristics of Big data. More recently, several other Vs have been added to description of big data, including veracity, value and variability.

1) Volume

The name Big Data itself is related to a size which is enormous. It is mainly about the relationship between size and processing capacity. Size of data plays very crucial role in determining value out of data and also, whether a particular data can actually be considered as a Big Data or not, is dependent upon volume of data. Therefore, volume is one characteristic which needs to be considered while dealing with Big Data.

2) Variety

Variety refers to heterogeneous sources and the nature of data. It describes the wide variety of data that is being stored and still needs to be processed and analyzed. Earlier spreadsheets and databases were the only sources of data considered by most of the applications. Now days, data in the form of audio and video files, photos, GPS data, medical files, instrument measurements, graphics, web documents etc.is also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

3) Velocity

The term velocity refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines the potential in the data Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks and social media sites, sensors, mobile devices, etc. The flow of data is massive and continuous.

4) Variability

This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

5) Value

The business value of the data collected and how big data gets better results from stored data. Big organizations use this characteristic to evaluate potential benefits.

6) Veracity

Veracity shows the quality and origin of data; it refers to the degree to which big data can be trusted. In short, it describes about the truth and authenticity of the data. In a sense, it is a hygiene factor.

3. Data Analytics Methodologies

Data analytics (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. The difference between data analysis and data analytics is that data analytics is a broader term of which data analysis forms a subcomponent. Data analysis refers to the process of compiling and analyzing data to support decision making, whereas data analytics also includes the tools and techniques use to do so. Data mining is defined as a process used to extract usable data from a larger set of any raw data. Sifting through very large amounts of data for useful information. Data mining is the process of extracting through large data sets to identify patterns and establish relationships to solve problems through data analysis. By "mining" large amounts of data, hidden information can be discovered and used for other purposes. Data mining involves six common classes of tasks. Anomaly detection, Association rule learning, Clustering, Classification, Regression, Summarization

A. Clustering

Clustering is the most powerful, popular and commonly used unsupervised data mining technique, which deals with grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities and finding groups in a set of unlabelled data. Basically, all clustering algorithms uses the distance measure method to predict the similar characteristics. In distance measure method, data points that are lying further in the data space exhibit less similar characteristics whereas the data points closer in the data space exhibit more similar characteristics. Different algorithms use different approach to find the similar characteristics. The clustering methods include Partitioning Clustering Method, Hierarchical Clustering Method, Density-Based Clustering Method, Grid-Based Clustering Method, Model-Based Clustering Method, Constraint-Based Clustering Method. Selecting appropriate clustering method and optimal number of clusters in healthcare can be quite difficult and confusing. On analysis K-means and DBSCAN algorithms seems to have strong inter-cluster separation and intra-cluster cohesion [7], but K-means algorithm have proven promising performance and opportunities to many aspects of healthcare practice.

1) K-Means Clustering Algorithm

In K-means clustering the data is divided into K clusters in such a way that each cluster contains at least one point and each point belongs to exactly one cluster. It uses simple iterative

technique to group the points in data set to form a cluster. Steps involved in K-means algorithm are:

1. Initially the number of clusters (K) is decided.
2. Assign each of the data point to the nearest centroid by calculating the minimum Euclidean distance of each data point to each centroid. The Euclidean distance between points a and b is given by,
   Dist (a,b) = Dist (b,a) =
3. Set the position of each cluster to the mean of all data points belonging to that cluster.
4. Repeat steps b and c until a convergence is observed.

B. Classification

Classification is a major technique in data mining and widely used in various fields. Classification is one of the data mining functions that assigns items in a collection to target categories or particular classes. Classification is used to accurately predict the target class for each case in the data. Types of classification algorithms include Linear Classifiers: Logistic Regression, Naive Bayes Classifier, Support Vector Machines, Decision Trees, Boosted Trees, Random Forest, Neural Networks, Nearest Neighbour, etc. Classification technique follows three approaches: Statistical, Machine Learning, Neural Network.

Akansha Priya and Er. Meenakshi [27] have proposed a C4.5 algorithm has been analysed using WEKA tool to find the phishing sites. C4.5 is an evolution of ID3. The C4.5 algorithm generates a decision tree for the given data by recursively splitting that data. A training dataset of 750 URLs has been made to train the algorithm The C4.5 algorithm allows pruning of the resulting decision trees. This increases the error rates on the training data, but importantly, decreases the error rates on the unseen testing data. Result shows C4.5 has an accuracy of 82.6%.

Concepción Burgosa, María L. Campanarioa, David de laPeñab , Juan A Lara , David Lizcanob , María A.Martínezb [28] have proposed the use of knowledge discovery techniques to analyse historical student course grade data in order to predict whether a student will drop out of a course or not. Logistic regression models are used for the purpose of classification. They conducted experiments with data on over 100 students for several distance learning courses confirm the predictive power of their proposal. Using the resulting predictive models, they have designed a tutoring action plan. Applying that plan, they have managed to reduce the dropout rate by 14% with respect to previous academic years in which no dropout prevention mechanism was applied.

Saptarsi Goswami, Sanjay Chakraborty, Sanhita Ghosh, Amlan Chakrabarti, Basabi Chakraborty [31] have focussed on reviewing the application of data mining and analytical techniques designed so far for prediction, detection, and development of appropriate disaster management strategy based on the collected data from disasters. Here a framework for building a disaster management database for India hosted on open source Big Data platform like Hadoop in a phased manner has been proposed.

C. Machine Learning

Humans have been learning, exploring a lot of knowledge in evolving technologies and producing in various sectors throughout the world, over the past few decades, Machine Learning has evolved exploiting of computer learning in mathematically considered computational approaches, comparing with enormous machine learning algorithms designed helping with commonly executable tasks more effectively. Machine Learning is a paradigm that may refer to learning from past experience (which in this case is previous data) to improve future performance. The sole focus of this field is automatic learning methods. In Data mining, is basically about interpreting any kind of data, but it lays the foundation for both artificial intelligence and machine learning. In practice, it not only sample information from various sources but it analyses and recognizes pattern and correlations that exists in that information that would have been difficult to interpret manually without today's advanced machine learning techniques comes in handy. Though numerous algorithms and techniques have been introduced as mentioned earlier, if it is closely studied most of the practical ML approach includes three main supervised algorithms are Naive Bayes, Support Vector Machine and Decision Tree. Majority of researchers have utilized the concept of these three, be it directly or with a boosting algorithm to enhance the efficiency further. The approaches used in machine learning as such as Supervised Learning, Unsupervised Learning, Semi-supervised Learning, and Reinforcement Learning like Regression Algorithm, Instance-based Algorithm, Regularization Algorithm, Decision Tree Algorithm, Bayesian Algorithms, Support Vector Machine (SVM), Clustering Algorithms, Association Rule Learning Algorithms, Artificial Neural network (ANN) Algorithms, Deep Learning Algorithms, Dimensionality Reduction Algorithms, Ensemble Algorithms. Supervised learning algorithms approximate the relation between features and labels by defining an estimator for a particular group of pre-labelled training data. The main challenge in this approach is pre-labelled data are not always readily available. So before applying Supervised Classification, data need to be pre-processed, filtered and labelled using unsupervised learning, feature extraction, dimensionality reduction etc.

Supanuth Ongsuk, Sakan Komolvatin, Intouch Kunakorntum, Phond Phunchongharn, Sumet Amonyingcharoen, and Woranich Hinthonghave [19] proposed in designing an adaptive modelling predicting machine learning prognosis on Cholangiocarcinoma which is a subset in liver cancer, which impacts Thailand in a major scale as in record. The study provides a prognosis framework as solution called "CanWiser" (adaptive) that is able to learn patients dataset containing demographic and laboratory test results, producing a recommendation model to reduce the risk of Cholangiocarcinoma based on the use of several classifiers calculating in considering sensitivity and specificity preferably rather than considering accuracy reasoning with producing high

misdetection, and to reduce risk of Cholangiocarcinoma based upon the use of k-means clustering model in achieving their goal.

Dhafar Hamed Abd, Jwan K. Alwan, Mohamed Ibrahim and Mohammad B. Naeem [24] have proposed a paper that focuses on the Sickle Cell Infection (SCI), which is a crucial for patients who are to be fully diagnosed in early stages with proper therapy procedures as fast as possible. This paper concentrates on expert systems providing physician's task through means of mobile-based platform system to facilitate the patients in managing and diagnosis towards enhancing to design a reliable home-based treatment recommendation system, for both non-critical and critical conditions based on the patient's sharing input data of their current analysis as blood and other tests data. As with the decision support system for non-critical conditions, and otherwise as for the critical outcomes are shared with human interacting doctor experts handling or attending the cases for further provisioning with treatment recommendations and suggestions. The idea behind to decease error rate as much as possible until machine learning can accommodate all possible dataset types in more efficient manner. Thus, their model proposed has an essential advantage in require of low storage and facilitate physician task by analyzing the huge amount of dataset accurately.

Marcia S Louis, Michael Alosco, Benjamin Rowland, Huijun Liao, Joseph Wang, Inga Koerte, Martha Shenton, Robert Stern, Ajay Joshi, Alexander P Lin [32] have proposed a paper that concentrates upon Chronic Traumatic Encephalopathy (CTE), a neurodegenerative disease due to trauma caused through head injuries. This paper designs a model for identification of the CTE with related biomarkers as a supervised machine learning problem. The authors proposed by designing two classification models, as to deduce the labels for supervised classifications, performing a univariate feature analysis using ANOVA algorithm attaining variance score against Metabolites to identify the important features necessary as following up with use of classifiers for making the desired classifications to understand the importance of each feature with KNN. As of the proposal, the study evident from the analysis that the neurochemical changes in brain corresponds with the neuropsychological test and clinical evaluation, thus demonstrating the strong value of machine learning techniques to evaluate changes in CTE.

Jiaxin Cai, Tingting Chen, Xuan Qiu [22] have proposed the diagnosis of fibrosis stage and inflammatory activity level in patients with chronic hepatitis C is important in clinical practices. To provide a non-invasive diagnosis for patients with chronic hepatitis C, this study proposed an automatic diagnosis system of chronic hepatitis C using serum indices data of patients to predict the fibrosis stage and inflammatory activity grade of chronic hepatitis C by training the extreme learning machine. They proposed an automatic diagnosis system is test on real clinical cases of chronic hepatitis C based on serum indices. Experimental results demonstrate that the performance of the proposed method exceeds that of the stateof-the-art baselines regarding the diagnosis of fibrosis stage and inflammatory activity grade of chronic hepatitis C.

Raid Lafta, Ji Zhang, Xiaohui Tao, Yan Li1, Xiaodong Zhu, Yonglong Luo and Fulong Chen [20] have proposed an effective medical recommendation system that uses a fast Fourier transformation-coupled machine learning ensemble model is proposed for short-term disease risk prediction to provide chronic heart disease patients with appropriate recommendations about the need to take a medical test or not a day advance based on analysing their medical data. The data are decomposed by use of the fast Fourier transformation in extracting the frequency information, a bagging-based ensemble model is utilized to predict the patient's condition the day in advance producing the final recommendation. Experimentally the proposed system yields a very good recommendation accuracy and offers an effective way to reduce the risk of incorrect recommendations. The results conclusively ascertain that the proposed system is a promising tool for analyzing time series medical data and providing accurate and reliable recommendations to patients suffering from chronic heart diseases

4. Tools used in Data Analytics

A. Apache Hadoop

Apache Hadoop is an open source software developed by Apache™ Hadoop® foundation for handling large volumes of data and large-scale processing of the data-sets on clusters of commodity hardware which is licensed under Apache License 2.0. The framework is used for achieving scalable, distributed and reliable computing. Hadoop is coded and released in java by Apache.

B. Hadoop Framework

Hadoop was founded by Doug Cutting and Mike Cafarella in 2005 which was developed to support distribution for the Nutch Search Engine project. The Hadoop is a source implementation of Map Reduce framework. Doug cutting inspired the idea of Hadoop framework from Google File System (GFS). ache Hadoop's MapReduce and HDFS components originally derived respectively from Google's MapReduce and Google File System (GFS) papers. The basic two layers of the Hadoop system are the Map reduce Engine, and the Hadoop Distributed File System (HDFS). The Map Reduce is a framework provides an environment to manage its Map execution and Reduce tasks across the cluster of machines. The Map Reduce works on the key concepts based on key and value pairs.

The Hadoop framework is an autonomic system; it implements the distributed data file systems. Hadoop distributes the big datasets for processing across set of computers as clusters, allowing each compute or machine provide its standalone storage and computations handled in a distributed manner rather depending on a centralized server. It uses the Hadoop Distributed File System (HDFS) is used as an

underlying layer in the implementation of Map Reduce rather than GFS.

The framework manages the distribution and execution of the tasks, collecting the outcome of the processed data and reporting the status to the user. The distribution of the tasks is done by decomposing (Input Splitting) the submitted job into set of map tasks, shuffles, sorts and framing the set of reduce tasks. the entire Apache Hadoop "platform" is now commonly considered to consist of a number of related projects as well: Apache Pig, Apache Hive, Apache HBase, and others.

C. HDFS Terminologies

The Hadoop distributed file system terminologies are Namenode, Datanode, DFS Client, Files/Directories, Replication, Blocks, Rack-awareness. The HDFS architecture stores the large volumes of data (in the range of gigabytes to terabytes) to be distributed across multiple machines as clusters. The Namenode acts as the mail metadata server. The HDFS file system has a secondary Namenode that regularly connects with the primary Namenode to build the snapshots of the primary Namenode without having to replay the entire journal of file-system actions. An advantage of using HDFS is data awareness between the job tracker and task tracker. The job tracker schedules map or reduce jobs to task trackers with an awareness of the data location. In Hadoop file system the File access can be achieved through the native Java API, the Thrift API, to generate a client in the language of the users' choosing (C++, Java, Python, PHP, Ruby, Erlang, Perl, Haskell, C#, Cocoa, Smalltalk, or OCaml), the command-line interface, or browsed through the HDFS-UI web app over HTTP.

D. Mapreduce Engine

The engine comprises of two important services as the Job Tracker and Task Trackers. The Job Tracker receives the MapReduce Jobs submitted by the client application, and furtherly assign (push) the work to available TaskTracker nodes present in the cluster, which its work is to keep the data as close as possible. With the help of the rack-awareness file system, the TaskTracker is aware which node contains the data, and which machines are nearby. Any part of failure will lead the respective part of the job rescheduled. The TaskTracker used on each node spawns-off a separate Java Virtual Machine (JVM) process in order to prevent the TaskTracker from failing, if the running job crashes the JVM. To keep in check of the jobs assigned status, the TaskTracker sends a heartbeat to the JobTracker.

E. Nextgen Mapreduce (YARN)

The MapReduce 2.0 (MRv2) or YARN revamped in the hadoop-0.23. The Apache™ Hadoop® YARN is a sub-project of Hadoop by the Apache that is introduced in Hadoop 2.0 that the resource management and processing components are separated. The YARN based architecture provides a more general processing platform that is not constrained to MapReduce.

The fundamental idea of the MRv2 is to split up the major two functionalities of JobTracker, resource management and job scheduling or monitoring, into separate daemons. In Hadoop 2.0, YARN handles the resource management operations that was handled by MapReduce, such as to be packaged new engines.

The major advantage of YARN in Hadoop 2.0 is enabling the capability to run multiple applications, where all sharing a common resource management. This advantageous capability allowed several industries/organizations to use YARN in building their applications to Natively Run IN Hadoop. YARN enhances Hadoop by obtaining scalability, Compatibility with MapReduce, Improved Cluster Utilization, Support for workloads other than MapReduce, Agility.

F. Apache Spark

Apache Spark is a fast cluster computing framework extending the Hadoop Modal to enhance various kinds of computing which includes Interactive Queries and Stream Processing. Apache Spark is the most actively developed open source project among data tools. The Spark framework is an optimized engines supporting general execution graphs as the Directed Acyclic Graph (DAG) Engine that optimizes the workflow. The Apache Spark has become the engine that to enhance many of the capabilities of the present Apache Hadoop Environment. For Big Data, Apache Spark meets a lot of needs and runs natively on Apache Hadoop's YARN. The Apache Spark is well integrated with the Apache Hive. By running Apache Spark in your Apache Hadoop environment, you gain all the security, governance, and scalability inherent to that platform. The engine provision working with the programming platforms as Java, Scala, Python and R, that works with MapReduce, HDFS, HBase, S3, MongoDB and so forth. The primary concept Resilient Distributed Dataset (RDD) that supports the way the Spark engine works. The main components are Spark core, Spark Structured Query Language (SQL), MLlib, Spark Streaming, and GraphX.

G. Resilient Distributed Dataset (RDD)

The RDD is a fundamental data structure of Spark, which is an immutable distributed collection of data or dataset. Each dataset of RDD can be computed on different nodes among the clusters. It facilitates two types of operations such as the Transformation and Actions. The transformation operations are such as filter(), map(), or union() on an RDD that yields another RDD. An action include the operation such as the count(), first(), take(n), or collect() that triggers a computation, that returns a value back to the Driver program which remembers the transformations applied to an RDD, or writes to a stable storage system like Apache Hadoop HDFS.

H. Hadoop Spark Machine Learning Library (MLLib)

The Spark engine became the hot choice which MLlib is a main reason providing a possibility of proving Machine Learning Algorithms Library. Apache Spark's MLlib 2.0 is an

open source, distributed, scalable, and platform independent machine learning library. The library specially designed to enabling simplicity, scalability, language compatibility, speed of processing (processing) and easy integration with several other data science tools. The Spark offers high-quality algorithms for classification, clustering, regression, recommendation, topic modelling, frequent item set, etc.

Spark enhances machine learning by addressing the problems faced by data scientists such as unlikely ready packages to help solve their data problems, handling movement of data becomes a time consuming that happens to be an extensive engineering at moving development to production environment. Spark runs on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud, against diverse data sources.

Spark Ecosystem – Spark Components

1. *Spark Core Component:* It is responsible for basic I/O functionality, scheduling and monitoring the jobs on spark clusters, task dispatching, networking with different storage system, fault recovery and efficient memory management.
2. *Spark SQL Component:* Leverages the power of declarative queries and optimized storage by running SQL like queries on Spark data that present in RDDs and other external sources.
3. *Spark Streaming:* Allows developers to perform batch processing and streaming of data with ease, in the same application.
4. *Spark MLlib:* MLlib eases the deployment and development of scalable machine learning pipelines.
5. *GraphX:* GraphX allows Data scientist to work with graph and non-graph sources to achieve flexibility and resilience in graph construction and transformation.

I. Apace Spark Tachyon In-Memory

Tachyon is a reliable shared memory that forms an integral part of the Spark ecosystem which helps to achieve the desired throughput and performance by avoiding unnecessary replications. The in-memory computation allows to perform interactive and fast queries. The Tachyon in-memory with increase in size of datasets and storage for different workloads supports reliable file sharing across computing frameworks as Spark and Hadoop at memory speed. Tachyon is already used in production at popular companies like RedHat, Yahoo, Intel, IBM, and Baidu. Tachyon code is contributed by 100+ developers from 30 organizations.

5. Conclusion

After reviewing upon the research articles towards medical

Table 1
Table of Survey and conclusions

Authors	Conference	Keywords	Comments
Supanuth Ongsuk et al.	1st IEEE International Conference on Knowledge Innovation and Invention 2018 [ongsuk2018]	Cholangiocarcinoma, Classification, Personalized Recommendation, Machine Learning	The study provides a prognosis framework that adaptively producing a recommendation model reducing the risk of Cholangiocarcinoma, considering sensitivity and specificity preferably rather than considering accuracy and to reduce risk of Cholangiocarcinoma.
DhafarHamed Abd et al.	Annual Conference on New Trends in Information and Communications Technology Applications(NTICT'2017)	Sickle Cell Disease; Machine Learning Algorithm; Mobile Healthcare Service; Real-time data; Self-care Management System; E-Health	The model proposed has an essential advantage in require of low storage and facilitate physician task by analysing the huge amount of dataset accurately.
Marcia S Louis et al.	2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)	Chronic Traumatic Encephalopathy; Machine Learning; Proton Magnetic Resonance Spectroscopy	The study evident from the analysis that the neurochemical changes in brain corresponds with the neuropsychological test and clinical evaluation, thus demonstrating the strong value of machine learning techniques to evaluate changes in CTE.
SaptarsiGoswami et al.	Ain Shams Engineering Journal Volume 9, Issue 3, September 2018, Pages 365-378	data mining, management strategy, Big Data, Hadoop	The authors proposed an application of data mining and analytical techniques designed so far for appropriate disaster management strategy.
Akansha Priya, Er. Meenakshi	2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017, India	Phishing sites, C4.5 (J48) data mining algorithm, WEKA tool, classification approach	The authors proposed C4.5 can handle both continuous and discrete attributes and can also deal with numeric attributes, missing values, and noisy data
Jiaxin Cai, Tingting Chen, Xuan Qiu	2018 9th International Conference on Information Technology in Medicine and Education [cai2018]	Chronic hepatitis C; fibrosis; inflammatory activity; machine learning; extreme learning machine	The proposed automatic diagnosis system Experimentally demonstrate the performance of the proposed method exceeds that of the state-of-the-art baselines regarding the diagnosis of fibrosis stage and inflammatory activity grade of chronichepatitis C.
Raid Lafta et al.	IEEE Access (Volume: 5)	Recommendation systems, time series analysis, intelligent systems	The results conclusively ascertain that the proposed system is a promising tool for analyzing time series medical data and providing accurate and reliable recommendations to patients suffering from chronic heart diseases

analysis and recommendation frameworks designed, worked with respective to cancers treatment regarding with brain damages, cell infections, and chronically medical attention predictions and recommendation generators. Thus, we have analyzed the future proposals, identifying their limitations of the work algorithms utilized along with used methodologies. Thus, gathering our interest upon the study in medical treatment recommendation techniques will be taken into implementation for prognosis upon predicting different disease and recommendation outcomes over collective data resources.

References

[1] C. Li , T. Chen , Q. He , K. Li , Mruninovo: an efficient tool for de novo peptide sequencing utilizing the Hadoop distributed computing framework, Bioinformatics 33 (6) (2016) 944–946.
[2] Erevelles S, Fukawa N, Swayne L (2016) Big data consumer analytics and the transformation of marketing. J Bus Res 69:897–904.
[3] Mohammed K. Hassan, Ali I. El Desouky, Sally M. Elghamrawy and Amany M.Sarhan: Big Data Challenges and Opportunities in Healthcare Informatics and Smart Hospitals.
[4] Apache, Hadoop, 2017, (Website). http://hadoop.apache.org .
[5] Chen , K. Li , Z. Tang , K. Bilal , K. Li , A parallel patient treatment time prediction algorithm and its applications in hospital queuing-recommendation in a big data environment, IEEE Access 4 (2016) 1767–1783 .
[6] Apache, Spark, 2017, (Website). http://spark-project.org .
[7] Godwin Ogbuabor and Ugwoke, F. N: Clustering algorithm for a healthcare dataset using silhouette score value, International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 2, April 2018.
[8] Alsayat, A., & El-Sayed, H. (2016). Efficient genetic K-Means clustering for health care knowledge discovery. In Software Engineering Research, Management and Applications (SERA), 2016 IEEE 14th International Conference on (pp. 45-52).
[9] Sadia Din, Awais Ahmad, Anand Paul, Muhammad Mazhar Ullah Rathore, Gwangill Jeon: A Cluster-Based Data Fusion Technique to Analyze Big Data in Wireless Multi-Sensor System, IEEE., Volume 5, 2017.
[10] Nur Al Hasan Haldar, Farrukh Aslam Khan, Aftab Ali and Haider Abbas, Arrhythmia Classification using Mahalanobis Distance based Improved Fuzzy C- Means Clustering for Mobile Health Monitoring Systems, Neurocomputing.
[11] Tiago Hillerman, Jo˜ao Carlos F.Souza, Ana Carla B.Reis, Rommel N.Carvalho, Applying clustering and AHP methods for evaluating suspect healthcare claims, Journal of Computational Science.
[12] KM Archana Patel and Prateek Thakral, The Best Clustering Algorithms in Data Mining, International Conference on Communication and Signal Processing, April 6- 8, 2016, India.
[13] Yinghua Lv, TinghuaiMa, MeiliTang, JieCao, YuanTian , Abdullah Al-Dhelaan , MznahAl-Rodhaan, An efficient and scalable density-based clustering algorithm for datasets with complex structures" Elsevier, 23 March 2015.
[14] Y. Anavi, I. Kogan, E. Gelbart, O. Geva, and H. Greenspan. A comparative study for chest radiograph image retrieval using binary texture and deep learning classification. Conf Proc IEEE Eng Med Biol Soc, 2015:2940–2943, Aug 2015.
[15] Ying Liu, Brent Logan, Ning Liu, Zhiyuan Xu, Jian Tang, and Yanzhi Wang, Deep Reinforcement Learning for Dynamic Treatment Regimes on Medical Registry Data, 2017 IEEE International Conference on Healthcare Informatics.
[16] Y. Wang, P. Wu, Y. Liu, C. Weng, and D. Zeng, Learning optimal individualized treatment rules from electronic health record data, International Conference on Healthcare Informatics (ICHI). IEEE, 2016, pp. 65–71.
[17] Mohammad-Parsa Hosseini, Tuyen X. Tran, Dario Pompili, Kost Elisevich, and Hamid Soltanian-Zadeh, Deep Learning with Edge Computing for Localization of Epileptogenicity using Multimodal rs-fMRI and EEG Big Data, 2017 IEEE International Conference on Autonomic Computing.
[18] M.-P. Hosseini, H. Soltanian-Zadeh, K. Elisevich, and D. Pompili, Cloud-based deep learning of big eeg data for epileptic seizure prediction, IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE, 2016.
[19] Supanuth Ongsuk, Sakan Komolvatin, Intouch Kunakorntum1, Phond Phunchongharn, Sumet Amonyingcharoen, and Woranich Hinthong, An Adaptive Cancer Prognosis Framework for Cholangiocarcinoma based on Machine Learning Techniques, IEEE International Conference on Knowledge Innovation and Invention 2018.
[20] Raid Lafta, Ji Zhang, Xiaohui Tao, Yan Li, Xiaodong Zhu, Yonglong Luo and Fulong Chen, Coupling a Fast Fourier Transformation with a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment.
[21] S. Li, B. Tang, and H. He, "An Imbalanced Learning based MDR-TB Early Warning System," Journal of medical systems, vol. 40, no. 7, pp.1-9, 2016.
[22] Jiaxin Cai, Tingting Chen, Xuan Qiu: Fibrosis and Inflammatory Activity Analysis of Chronic Hepatitis C Based on Extreme Learning Machine, 2018 9th International Conference on Information Technology in Medicine and Education.
[23] J. Cai, T. Chen, Y. Li, N. Zhu, and X. Qiu, "A novel collaborative representation and scad based classification method for fibrosis and inflammatory activity analysis of chronic hepatitis c," in Young Scientists Forum, 2018.
[24] Dhafar Hamed Abd, Jwan K.Alwan, Mohamed Ibrahim, Mohammad B. Naeem, The Utilisation of Machine Learning Approaches for Medical Data Classification and Personal Care System Management for Sickle Cell Disease, Annual Conference on New Trends in Information; Communications Technology Applications (NTICT'2017) -9 March 2017.
[25] Forkan A, Khalil I, Ibaida A, Tari Z (2015), "BDCaM: big data for context-aware monitoring a Personalized knowledge discovery framework for assisted healthcare" IEEE Trans Cloud Comput. pp 1–1.
[26] J. Xie, H. Gao, W. Xie, X. Liu , P.W. Grant, "Robust clustering by detecting density peaks and assigning points based on fuzzy weighted k-nearest neighbors," Inf. Sci. 354 (2016), 19–40 .
[27] Akansha Priya, Er. Meenakshi, Detection of Phishing Websites Using C4.5 Data Mining Algorithm, 2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017, India.
[28] Concepción Burgosa, María L. Campanarioa , David de laPeñab , Juan A Lara , David Lizcanob , María A.Martínezb have proposed the use of knowledge discovery techniques in 2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017, India.
[29] Kajaree Das, Rabi Narayan Behera, "A Survey on Machine Learning: Concept, Algorithms and Applications," in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017.
[30] Somaya Hashem, Gamal Esmat, Wafaa Elakel, Shahira Habashy, Safaa Abdel Raouf, Mohamed Elhefnawi, Mohamed, El-Adawy, and Mahmoud ElHefnawi, "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients," in IEEE/ACM Transactions on Computational Biology and Bioinformatics.
[31] Saptarsi Goswami, Sanjay Chakraborty, Sanhita Ghosh, Amlan Chakrabarti,and Basabi Chakraborty, "A review on application of data mining techniques to combat natural disasters," Ain Shams Engineering Journal.
[32] Marcia S Louis, Michael Alosco, Benjamin Rowland, Huijun Liao, Joseph Wang, Inga Koerte, Martha Shenton, Robert Stern, Ajay Joshi, and Alexander P Li, "Using Machine Learning techniques for identification of Chronic Traumatic Encephalopathy related spectroscopic Biomarker," IEEE Applied Imagery Pattern Recognition, 2017.