Machine Learning and Big Data Implementation on Health Care data

Gopinadh Sasubilli,
Big Data Application Specialist, Merck,USA
gopinadh.sasubilli@merck.com

Abhishek Kumar,
SMIEEE, Chitkara University Institute of Engineering, and technology, Chitkara University, Punjab, India.
Abhishek.kumar@chitkara.edu.in

Abstract: **Healthcare is the most prominent field suitable for the applications of machine learning and big data on health care data. The implementations of health care with big data and machine learning is increased with the client health requirements. The electronic health record applications are being increased in this current situation, which is needed to be focused on utilizing the data generated by those applications. There is a large volume of data in health care that is related to different health care domains especially neuro and cardiac. These data need a special focus and the architectures currently focusing on these domains has to implement the latest technologies to predict some patterns. In this article, the implementation of different health care architecture is focussed, which uses live data gathered from different sources over the globe. In this article, machine learning approaches and the big data framework are combined to design a prediction model and data handling techniques.**

Keywords: Big Data, Medical, Health care, Prediction, Data

I. INTRODUCTION

Big data analytics plays a crucial role in managing the prediction models and the data acquired from different sources make security messes and architecture failures to manage the live data. The different kinds of data models are being explained with the implementation of the prediction model in the medical stream is being explained in this article Healthcare industry is the large composition of different sub-concepts and all the domains will share the common problem of

gathering maybe an MRI or X-Ray which explains the disease of a patient and all the information will be managed by the admin in a repository and

consistency of the data and the genuineness of the data collected. How the gathered data is trusted and processed are the main approach needed to focus on implementing any kind of prediction model. As a data scientist, gathering and managing the data from different resources and the pre-processing of the information and the reusing the same information for different purposes is a challenging task. The task focussed in this article is in two stages. First stage deals with the survey on different approaches currently handled with the medical domain, concerning the big data. In the second approach, dealt with explaining the different architectures currently in big data framework for handling the medical information. Also, focus on the proposed architecture which is showing a better result on handling the data from different global resources. The analysis of huge big data is a major problem, the world is facing and the current data gathered in a day is in petabytes and in which 75% of the data belongs to the medical domain. The various problems and scenarios in the medical domain mostly with big data are to be focussed and understood. The big data, four V's system, concerning the medical domain is explained as follows [1]

i. Volume:

A huge amount of data is created in the medical domain which surpasses all the kinds of data created from any other domain. Research in the health care industry and DNA research creates more data every second in the form of text and image[2].

In this scenario, how this kind of large data is being created in a short time and also the type of data to be considered. The specific problem like COVID - 19 has no background information and needed to take the samples of the DNA from the positive cases and have to make the research with different compositions and this will create more amount of data which will be one pattern. Likewise, different

978-1-7281-4876-2/20/$31.00 ©2020 IEEE

patterns in the DNA and the different researches carried out in our daily life[3].

ii. Velocity:

How speed the data is being created is most important to understand to manage the framework and make it ready to accept this kind of large flow of data. This kind of large flow of data can make some traffic congestion. The main implementation lies in the framework being used for data management. Different tools in big data can handle this kind of related issue.

iii. Variety:

What kind of data gathered is the most important. In this scenario, the information gathered from the trusted resources and the type of data is also important. There are three kinds of data which are mentioned below.

a. Structured Data

This is the systemic format and this can be in the form of a pre-designed structure with rows and columns of the data. The data which can be used for the prediction mostly will be in the form of text, in a semantic format and this can be used for the prediction of something concerning the problem.

b. Unstructured Data

The unstructured data is in the format of multimedia and that can be in different formats like videos, audio, and images. These data are got from different sources in recent researches. Like MRI scans for brain diseases, have different scenarios to gather the unstructured data like multimedia for analyses.

Semi-Structured Data

The semi-structured information is in the combination of both structured and unstructured. In this scenario, XML or JSON are used as the main data gathering formats. In this approach, how to gather different types of information at the same time without spoiling the architecture of the data pattern is focussed.

iv. Veracity

Trustworthy of the data gathered from different sources is discussed. All the data cannot use all the information gathered and cannot perform trial and error methods for this kind of live data. The decision models will be spoiled if there is any kind of mismatch or not required data.

There is a standard survey on the healthcare implementations in the big data using the three kings od predictions like predictive modeling, description modeling, and prescriptive modeling. Figure 1 states that the approach from the survey [1].

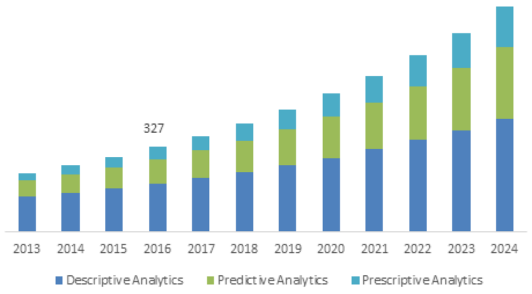

Figure 1: German Analysis of growth big data in health care

There are different kinds of sources and discuss them in detail of the existing systems in the literature survey. The lateral part of the article deals with the following things related to the big data with a literature survey, existing system, modeling different domains, proposed system, and conclude with the best approach in the existing systems.

II. LITERATURE REVIEW

[1] discusses the big data approach related to personalized health care management using big data. In this architecture author focused on the virtual physiological human (VPH) which deals with communicating with the doctor through the device which is our assistance. This article discussed the 5th V of the big data. That is value. Importance of the big data implementation in health care and personalized information is mentioned in this article. In this article author mentioned the procedure they processed a way to attach doctors and the engineers to provide a path for the better implementation of personalized health care services.

[2] Deals with the implementation of the electronic health records (EHR) which is a personalized health application through watch or mobile application the health of the patient will be monitored and the implementation deals with the requirements like what are the personal health information of the patient. In 2009 clinical standards have been changed by the WHO and the implementation of the medical stream changed with a lot of modifications within the system concerning the security and the data manipulation. Figure 2 explains the implementation of [2]

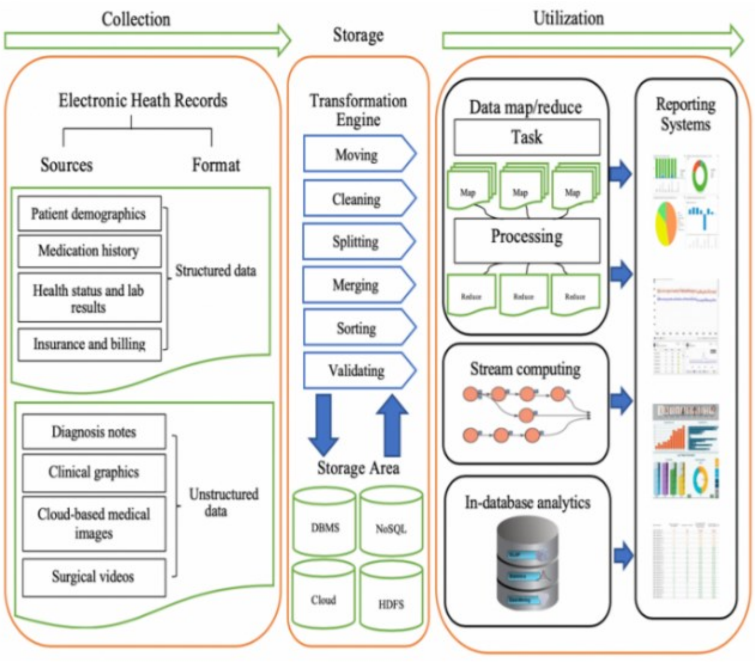

Figure 2: Health care records with Big Data Analytics

[3]deals with the implementation of the internet of things (IoT) with the health care industry. In this model, some identifications of low risk and moderate risk and high-risk management. 80% of medical services are allotted to the cities rather than the rural areas. Here in this article, an architecture is proposed that can carry out the prediction of whether the disease in the low risk or high risk. So that got clarity on cure or prevention with the respective treatment. Figure 3 explains that architecture. In this architecture, there will a relationship between the health condition captured using the IoT device and according to the prediction algorithms needed to understand whether they are low or high risk. So that whether to prevent or cure the problem.[4] deals with different approaches to the prediction analysis. The three different types of analysis of the data. 1) Predictive analytics which explains the outcome of the problem with a specific solution based on the previous experience, concerning the algorithm. 2) descriptive analytics deals with the concept of explaining the problem and the solutions for the related problems based on the requirement of the problem definition. 3) Prescriptive analytics deals with the implementation of the detail's insight into the problem with the root cause and deals with the previous experience in the same domain of the problem. Figure 4 deals with the pathways that have an implementation in big data

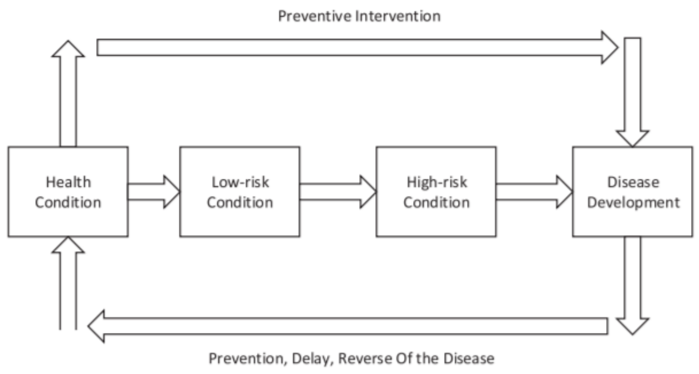

Figure 3: Health Management Model

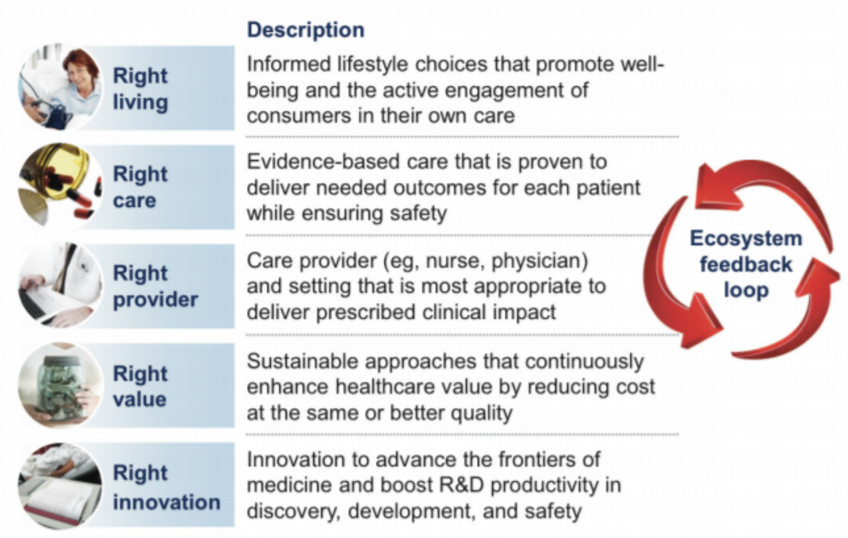
Description
Right living
Informed lifestyle choices that promote well-being and the active engagement of consumers in their own care
Right care
Evidence-based care that is proven to deliver needed outcomes for each patient while ensuring safety
Right provider
Care provider (eg, nurse, physician) and setting that is most appropriate to deliver prescribed clinical impact
Right value
Sustainable approaches that continuously enhance healthcare value by reducing cost at the same or better quality
Right innovation
Innovation to advance the frontiers of medicine and boost R&D productivity in discovery, development, and safety
Ecosystem feedback loop

Figure 4: Pathway of health care to big data

[5] deals with identifying the human activity patterns using big data and these patterns define the process of understanding the human emotions and behavior to identify the health patterns of the human

III. EXISTING SYSTEM

The existing system does not focus on the implementation of the cloud domain and live streaming management. Different mechanisms like Spark with Mahout Machine learning for the data storage and the prediction model design is concerned in the cloud services. But the live streaming which is the main problem has not discussed until now in the different articles. The live streaming concerning the medical domain is not yet discussed and implemented. In this current existing system, data management with constant size and weight is there, but there is no source of manging with the live stream. The lateral part of this section will discuss the live streaming mechanism concerning the medical domain.

IV. PROPOSED SYSTEM

The proposed system discusses two different kinds to maintain the best method to manage health care with security and not vulnerable. The two parts of the proposed approach are mentioned below. This article is designed, according to the below scenarios.

i. Best Approach to the current methods

The best methods cant be identified from the creation of one architecture [11-12]. All the architectures are not designed for the same kind. Figure 2 is the best kind of implementation until now. In that architecture, different kinds of implementations of health care and approaches to handle the problem [6-8] are shown.

There is a separation in the data collection and refinement process and also there is a separate framework and the process for the storage mechanism. In this storage mechanism, have different approaches. Cloud sources or any other kind of private storage mechanism can be used. And in the utilization of the data, using the MapReduce part which will separate the data into the chunks and those will be remapped according to the feature type or the category type.

ii. The Proposed approach.

The Spark, as big data framework, used for the implementation of the live data. The processing of live data is the most important issues in the big data and this can be a very useful implementation for the existing methodology. Kafka is the best method for the management of the better implementation of live streaming data [9-10]. The live streaming data like stock exchange and different medical data. For example, flight data can also be live data needed to manage. [13-14]. Streaming of the Spark was explained in figure 5 streaming.

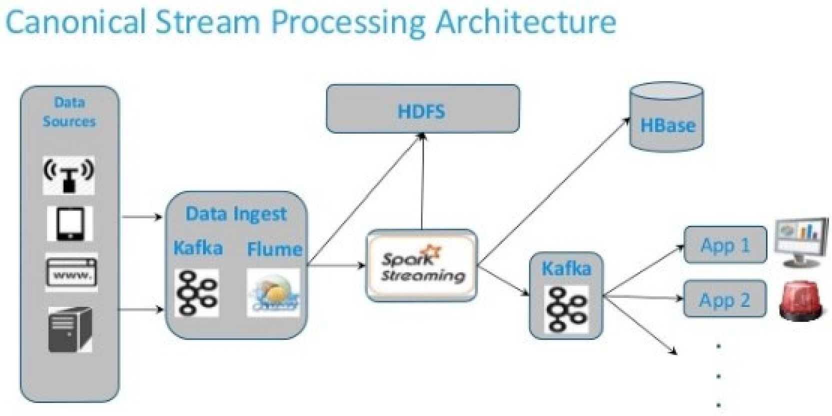

Figure 5: Spark streaming for live data management

Though it was working on the spark, Kafka and flume for the data ingest mechanism is also reliable [15]. To be more precise, if the patient is having some issue and if there is no chance of asking for help or informing to the doctor of the caretaker, then using the machine learning algorithm, an IoT device which is connected to the cloud will take the details of the health of the patient and the information which was considered as the input will be verified by the machine learning models as that is genuine or not[16-17]. If the data is genuine then the information will be processed and the doctor will get the information of the patient. In this regard, the data is live data. Using the live data and with the experience, the device has inbuilt with the machine learning models. It will implement some internal mechanisms and audit the live data and handle this kind of live data. The scala as the main framework and Kafka as the data ingestion from a big data platform.

V. CONCLUSION

Big data is the most important mechanism in recent times which was designed to manage the data in different forms. All the information, gathered from different resources will be pre-processed using the data mining mechanisms and those data features are forwarded to the frameworks according to the implementation and the problems. In this article, the medical implementation of different big data platforms is discussed. The information gathered from different resources will subject to the prediction models and the information gathered must be genuine and there must be no data lose in it. If the feature was built on the big data platform using any kind of cloud storage then needed to map the same to the repository after the manipulation of the result. Initially, the health care data will be gathered from different repositories and the storage mechanisms will monitor the data and store according to the file format. The saved information then utilized according to the requirement.

References

[1] “Big Data, Big Knowledge: Big Data for Personalized Healthcare” by Marco Viceconti, Peter Hunter, and Rod Hose, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19, NO. 4, JULY 2015

[2] “Optimizing the Electronic Health Records Through Big Data Analytics: A Knowledge-Based View” CAIFENG ZHANG , RUI MA , SHIWEI SUN, YUJIE LI , YICHUAN WANG , AND ZHIJUN YAN, VOLUME 7, 2019

[3] “Big Health Application System based on Health Internet of Things and Big Data” YUJUN MA , (Member, IEEE), YULEI WANG , JUN YANG , YIMING MIAO , AND WEI LI, Volume 5, 2017, Special section on HealthCare Big Data.

[4] “Big Data Analytics for Healthcare Industry: Impact, Applications, and Tools” Sunil Kumar and Maninder Singh, Volume 2, Number 1, Page No: 48-57,Big data mining and analytics.

[5] “Mining Human Activity Patterns From Smart Home Big Data for Health Care Applications” ABDULSALAM YASSINE ,(Member, IEEE), SHAILENDRA SINGH , AND ATIF ALAMRI, (Member, IEEE), Volume 5, 2017, special section on advances of multisensory services and technologies for healthcare in smart cities

[6] “Analysis Method of Motion Information Driven by Medical Big Data” JINGYI ZHANG et al. Volume 7, 2019, special section on data-enabled intelligence for digital health

[7] “Health Big Data Analytics: A Technology Survey” GASPARD HARERIMANA et al, Volume 6, 2018,

[8] “Big Data Visualization in Cardiology—A Systematic Review and Future Directions” SHAH NAZIR et al, Volume 7, 2019

[9] “In Search of Big Medical Data Integration Solutions - A Comprehensive Survey” HOUSSEIN DHAYNE et al, Volume 7,2019

[10] “Radiogenomics for Precision Medicine With a Big Data Analytics Perspective” Andreas S. Panayides et al, Volume 23, No:5, September 2019

[11]“Intelligent Analysis of Medical Big Data Based on Deep Learning” by HANQING SUN et al, Volume 7, 2019, special section on deep learning algorithms for internet of medical things

[12]"Health Big Data Classification Using Improved Radial Basis Function Neural Network and Nearest Neighbor Propagation Algorithm" by CONGSHI JIANG et al, Volume 7, 2019,special section on data-enabled intelligence for digital health

[13] "Analyzing Healthcare Big Data With Prediction for Future Health Condition" by PRASAN KUMAR SAHOO el al, Volume 4, 2016

[14] "Proposal of a health care network based on big data analytics for PDs" by Leonarda Carnimeo et al, Volume 4, Issue 6, The Journal of Engineering

[15] "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial" by HAN HU et al, Volume 2, 2014

[16] "The Internet of Things for Health Care: A Comprehensive Survey" by S. M. RIAZUL ISLAM et al, Volume 3, 2015

[17] "Harnessing the Power of Machine Learning in Dementia Informatics Research: Issues, Opportunities, and Challenges" by Gavin Tsang et al, Volume 13, 2020

[18] "Clinical big data and deep learning: Applications, challenges, and future outlooks " by Ying Yu et al, Volume 2, 2019, Issue 4