Novel Big Data Approach for Drug Prediction in Health Care Systems

Mamoon Rashid,
School of Computer Science and Engineering
Lovely Professional University
Jalandhar, India
mamoon873@gmail.com

Mir Mohammad Yousuf
School of Computer Science and Engineering
Lovely Professional University
Jalandhar, India
miryusufmir@gmail.com

Balwant Ram
School of Computer Science and Engineering
Lovely Professional University
Jalandhar, India
balwantramjassal@gmail.com

Vishal Goyal
Department of Computer Science
Punjabi University
Patiala, India
vishal.pup@gmail.com

***Abstract*— In Health Care Systems, consuming of medicines has become day to day activities for the people who are suffering from diseases. Most of the people are not also aware of the medication prescribed by doctors or pharmacies. Sometimes patients get other kind of complications as well by taking the medicines prescribed by medical practitioners. To counter these challenges, the authors are proposing the drug prediction model which will help patients for taking right medicines for the cure of particular disease. MLLib Library of Apache Spark is to be used for initial data analysis for drug suggestions related to symptoms gathered from particular user. The model will analyze the previous history of patients for any side effects of the drug to be recommended and considers weather and maps API from Google as well so that the patients can easily locate the nearby stores where the medicines will be available.**

***Keywords*—Health Care, HDFS, Apache Spark, Machine Learning, MLlib Library, Drug Discovery.**

I. Introduction

Machine Learning is presently playing major role in Health Care Systems by using various forms of data accumulated over years to derive meaningful insights. Health Care Systems are actively making use of machine learning along with Big Data Analytics to provide proper diagnosis and solutions for diseases by predicting right kinds of drugs. Whenever patient's complaint for any kind of disease, all symptoms are recorded and forwarded to computer with machine learning intelligence. Physicians usually recommend patients to undergo various tests and the inferences are carried out to resolve patient problems by using machine learning approach. For example, once the patient visits any consulting physician, the next step is to take scans in terms of X-rays and MRI's. These scans are later provided as input to machine learning models to diagnose patient problems and health condition with better results.

The inclusion of Big Data Analytics has brought new opportunities for treating patients in the domain of drug development and precision medicines. The use of Big Data Analytics along with Machine Learning has transformed health care systems to the next level. However still Health Care systems are fighting for the right understanding of diseases and drugs. According to Schork, N. J. et al. in [1], only 25% patients are benefitted from the top 10% drugs which are prescribed in United States Health Care Systems. This percentage is only 2% for patients who are prescribed for cholesterol drugs.

The implementation of Big Data helps in tasks for maintaining data in terms of Electronic Health Records and brings data in perfect shape for data monitoring. Big Data is playing its vital role for bringing global medical system together and allowing places and countries to get best treatments and consultation. Social media is the medium where Big Data Analytics has contributed in Health Care Systems. People speak about diseases on social networking sites like Facebook and Twitter.

This kind of real data is to be analyzed for insights for various kinds of health care information's by various Big Data Techniques and help in awareness among masses at global level [2]. The valuable insights can be drawn out of clinical data by the use of smart healthcare technology in terms of big data analytics.

This process achieves success in presenting patients risk forecast. This approach will certainly replace the expensive procedures used for maintaining records for patients in Health Care Systems [20]. Big Data Technology has allowed to store huge amounts of patient data in terms of quantity and thus to continuously analyze it for improving quality of life. Big Data Market in terms of Health Care Systems is estimated to grow its market place from 10 billion dollars in year 2016 to 27.6 billion dollars by year 2021 [3].

The outline of the paper is structured as follows: Section II gives the background of Drug Prediction work in Big Data Analytics. Section III proposed the model based on Big Data pipeline for effective prediction of drugs. In Section IV, the possible outcomes of this proposed model are given where the model will prove effective. Conclusion and future scope of the approach is given in section V.

II. Background

Vangsted, A. J et al. in [4] developed drug response prediction model for gene expression profiling from tumor samples. The authors in this work identified the patients with myeloma having high sensitivity for drugs for various suffering toxicity. A machine learning approach along with feature selection technique is performed for the analysis of peptides. This work has given SMO based classifier which predicted the presence of lantibiotics with an accuracy of 88.5% [5]. The application of artificial intelligence in health care systems was discussed in [6] for past, present and future.

The work is outlined to use Artificial Intelligence for cardiology, neurology and cancer. This work has provided the detailed review for the detection and treatment of these diseases in health care systems. The real time processing of health care data has been projected in [7] where data from different medical related applications and mobile applications stored in Electronic Medical Records is brought in to hadoop and MongoDB environments. This work approach minimized the processing time in patient records to greater extent. Dimitri et al in [8] have devised machine learning algorithm, DrugClust, which predicts the side effects in drugs. In this research, the devised machine learning algorithm first clusters the drugs on the basis of their features and then later Bayesian scores is used to predict the various side effects of drugs.

The results achieved in this research are promising when evaluated by using 5-folds cross validation procedure. The computational method was proposed in [9] for the prediction of drug-drug interactions. This model claims predictions of 250,000 unknown drug-drug interactions. Predictions in this model are based on similarities in drugs which are functional in nature. Harnie, D. et al. in [10] used apache spark based pipeline for scaling target predictions in drugs using machine learning approach. The authors in this research claim a speedup factor up to 8 nodes linear in nature and thus enhancing the processing performance in drug discovery. This work basically partitioned the work among various compounds and provided intermediate results. The network bandwidth and time is saved by processing the intermediate data on the same nodes that produce it.

Lo, Y. C. et al. in [11] have provided machine learning approach for mining chemical information from chemical databases for drug discovery. This research has provided a means of extracting and processing data related to chemical structures for identifying drugs with important biological properties. The work is tested for various machine learning models and utility of each model is discussed as well. Chen, R. et al. in [12] have provided a detailed review of machine learning for drug target interaction prediction. This research has highlighted the various databases which are used for drug discovery. The various classification schemes are outlined as well with methods for each category.

This research further discusses all the challenges of machine learning for drug target predictions. Panteleev, J. et al. in [13] discussed the recent advancements in machine learning for drug discovery. This research outlined approaches in deep learning for synthesis and design of compounds, binding predictions and other important properties. This research concludes that machine learning aims to reduce cost, labor demands and cycle time in early levels of drug discovery. The study in [14] have provided the perspective of machine learning to that of medical education. The authors put emphasis on machine learning inclusion among medical students, residents and fellows. This research directs educational systems especially medical side to include machine learning in curricular time and draw valuable insights. Lavecchia, A. in [15] has discussed the various methods and applications based on machine learning for discovery of drugs. The major focus given in this research is given to machine learning techniques for ligand based virtual screening.

The limitations have been discussed in detail with opportunities and successes kept under consideration as well. Zhang, L. in [16] summarized various deep learning approaches of machine learning along with applications for discovery of rational drugs. This research suggests that big data pipeline along with machine intelligence can be helpful guide for design and discovery of drugs. The research in [17] discusses the remarkable achievements of drug discovery with the use of deep learning procedures.

This research concludes that deep learning is having more flexibility for its architecture in comparison to machine learning and there is ease for creating neural network architectures in deep learning. However this research outlines the limitation of deep learning for its need of large training sets.

The authors in [18] has discussed machine learning and statistical techniques and approaches for the prediction of protein-ligand interactions in discovery of drugs. This research outlined the major difficulties and challenges for the prediction of ligands and highlighted challenges for using datasets which are unbiased for models.

The study in [19] have given overview of using big data in drug discovery. This research concludes that artificial intelligence with NLP pipelines can draw successes in the field of drug discovery. Moreover the authors have given outline to prepare models in Big Data for handling various issues in discovery of drugs.

III. PROPOSED APPROACH FOR DRUG PRIDICTION

Nowadays medicine consuming has become day to day activities for the people who are suffering from diseases. Most of the people are not also aware of the medication prescribed by doctors or pharmacies. Sometimes patients get other kind of complications as well by taking the medicines prescribed by medical practitioners. There are many reasons associated with it, according to Forbes, some problems associated with it are here.

Overuse and unnecessary care accounts for a high amount of money and is more common than you might imagine. Unnecessary tests and drugs explain Why Health Care Costs So Much.

1. Traditionally, health plans, Medicare and Medicaid pay providers for whatever services they deliver, regardless of whether services truly benefit the patient.

2. Transparency galvanizes change like nothing else. Transparency is a vital component to build an effective and efficient health care system and the lack of transparency in Indian healthcare threatens to erode public trust.

3. Awareness of the people: A lot of primary health problems can be solved if we provide effective training and the knowledge to the local population. The lack of awareness among the patients provides doctors to take benefit of "supplier induced demand" to extract money.

4. Accessibility: The rural-urban divide is enormous; therefore, proper supply chain management is indispensable. The usage technology is good for accessibility, the use of telemedicine is very helpful.

For addressing the above challenges, the authors tried to eradicate these issues somehow by prescribing the optimal drug, its usage and its availability by giving the location of the retailer of that drug near the customer. The proposal of big data pipeline along with machine learning approach for the prediction and suggestion of right drugs is shown in Figure 1.

The idea is to prepare Big Data Pipeline and store Health Care Data in terms of patient's previous summary and medicines in Hadoop Distributed File System. Later when the patient will come with any kind of symptoms, machine learning algorithms are applied in terms of Mllib library of Apache Spark to provide drug suggestions keeping side effects of patient under consideration on the basis of previous log data. The machine learning approach is to be used iteratively until the best suited medicine is to be suggested for patient. Later concept of GPS API is to be added in idea which will provide the availability of location where the medicine is to be available. The various functions performed by this big data pipeline is explained in steps:

A. Drug Analysis and Prediction

Providing the transparency to the people who do not know that which medicine is good for the disease and which is not. Saving the patients from the wrong diagnosis by the doctors. Based on the historical datasets and machine learning algorithms, optimal drug is prescribed according to the disease input by the user.

B. Side Effects Analysis

This module is for analyzing patient's data that can lead to side effects by that prescribed drug. A medicine may have a side-effect of skin-irritation. But if the patient is already having some other skin problem then taking this medicine can worsen the condition of patient.

C. Next Optimal Drug Suggestion

This module follows the above module. If the prescribed drug leads to side effects based on the drug records and patient records, then the drug dataset is analyzed again and next optimal drug is suggested.

D. GPS API

Geo location refers to the identification of the geographic location of a user or computing device via a variety of data collection mechanisms. Typically, most geo location services use network routing addresses or internal GPS devices to determine this location. By adding the GPS API, we will be able to fetch the location of the user.

E. Drug Location Availability

Based on the user's location, drug retailers and medicinal shops near the user's location are shown to the customers.

F. Speech Recognition

Speech recognition will be implemented for the blind and the physically disabled so that they can speak the disease name and they would be suggested for the medicines which can be an antidote to their disease, especially useful for users with disability.

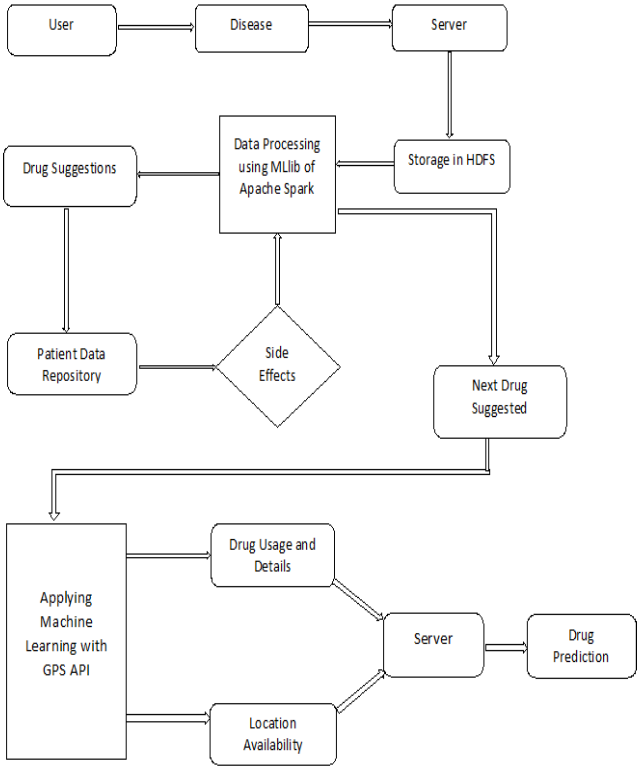

Fig. 1. Proposed Drug Prediction Model using Big Data and Machine Learning

The drug details and usage will also be given to patient for effective use of drugs. The step procedure of drug prediction model using big data pipeline is shown in Figure 2.

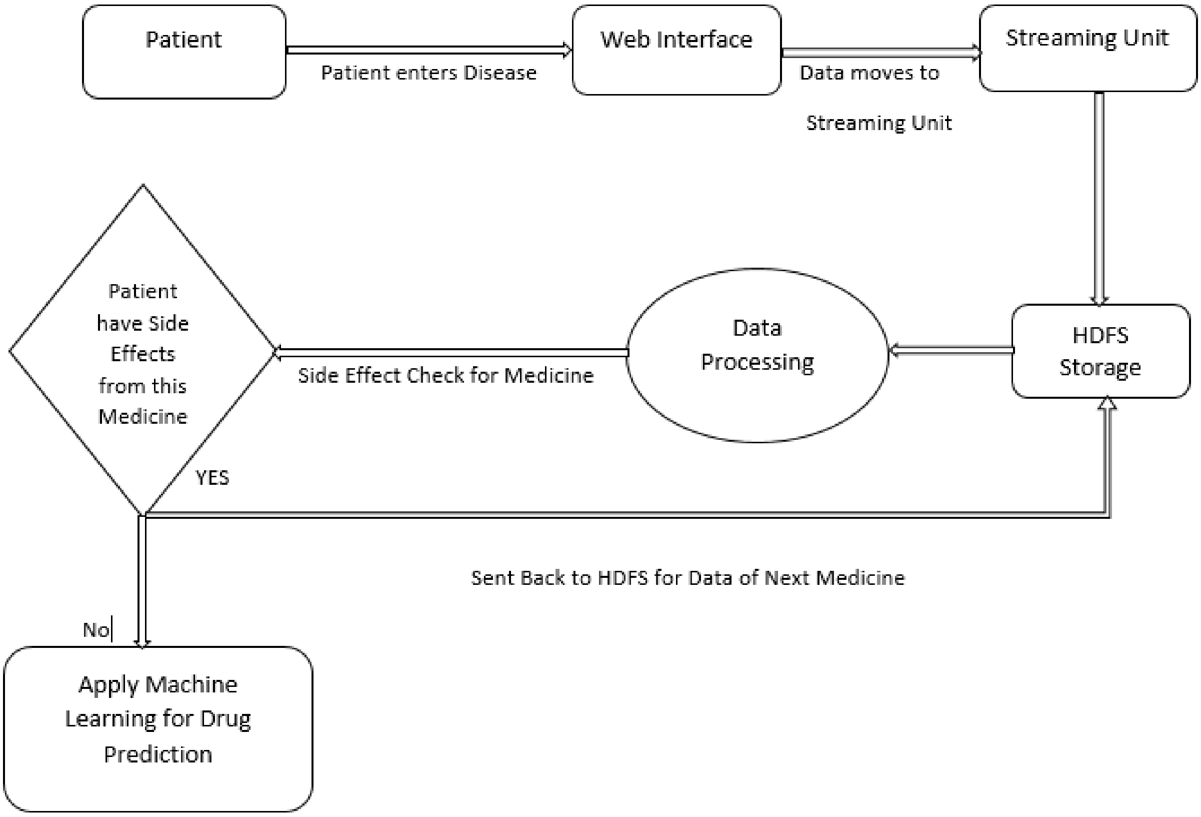
Patient
Patient enters Disease
Web Interface
Data moves to
Streaming Unit
Streaming Unit
HDFS Storage
Data Processing
Side Effect Check for Medicine
Patient have Side Effects from this Medicine
YES
Sent Back to HDFS for Data of Next Medicine
No
Apply Machine Learning for Drug Prediction

Fig. 2. Step Procedure of Drug Prediction Model using Big Data Pipeline

The patient is required to enter disease name or symptoms to the web interface from where the data is to be fetched and stored in Hadoop Distributed File System (HDFS). Patient query is checked from the log data present on HDFS and machine learning approach is to be applied to check side effects of drug suggested. Alternate medicines are suggested in case the recommended drug is having side effects for patient.

IV. EXPECTED OUTCOME OF WORK

A. Save Time and Money

Going to hospital costs a great deal of time and money. You have to make an appointment in advance but sometimes still wait for hours to see doctors, and then stand in line to get your medicines. By utilizing data analytics in this model, last health records can be used to reduce diagnostic process and avoid cost-intensive treatments that they did not work in history.

B. Better Care

This idea will help the patients to get confirmation that they are being diagnosed for the correct disease by entering the symptoms and then getting the disease name. Now, the second problem may be that doctor have identified the diseases correctly, but medicine is not proper for it or it has some side-effects. This problem will also get solved by getting the list of medicines from HDFS for that disease. Third thing is that sometimes patients find it difficult for getting a particular medicine because it may not be available on local stores. So, authors are proposing the availability of locations of various medical stores where that medicine can be found.

V. CONCLUSION AND FUTURE WORK

This study of research will surely eradicate the issues somehow by prescribing the optimal drug, its usage and its availability by giving the location of the retailer of that drug near the customer. This research chapter takes a step-in order to reduce various errors in medical system. It will reduce the unnecessary costs and overuse of drug which is more common in today’s world. It will also bring the transparency, along with awareness regarding the details of medicines and its usage which is a vital component to build an effective and efficient health care system. The configuration of multi node cluster with Apache Spark will further make this framework robust and enhance computations for lesser processing time.

REFERENCES

[1] Schork, Nicholas J. "Personalized medicine: time for one-person trials." Nature 520, no. 7549 (2015): 609-611.

[2] Bachrach, Yoram, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. "Personality and patterns of Facebook usage." In Proceedings of the 4th annual ACM web science conference, pp. 24-32. ACM, 2012.

[3] Kalyan Banga, “Big Data Healthcare Market to reach $27.6bn by 2021. Future Analytics World”, October, 2016. Retrieved from http://fusionanalyticsworld.com/big-data-healthcare-market-reach-27-6bn-2021/

[4] Vangsted, A. J., S. Helm-Petersen, J. B. Cowland, P. B. Jensen, P. Gimsing, B. Barlogie, and S. Knudsen. "Drug response prediction in high-risk multiple myeloma." Gene 644 (2018): 80-86.

[5] Poorinmohammad, Naghmeh, Javad Hamedi, and Mohammad Hossein Abbaspour Motlagh Moghaddam. "Sequence-based analysis and prediction of lantibiotics: A machine learning approach." Computational biology and chemistry 77 (2018): 199-206.

[6] Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. "Artificial intelligence in healthcare: past, present and future." Stroke and vascular neurology 2, no. 4 (2017): 230-243.

[7] Basco, J. Antony, and N. C. Senthil kumar. "Real-time analysis of healthcare using big data analytics." In IOP Conference Series: Materials Science and Engineering, vol. 263, no. 4, p. 042056. IOP Publishing, 2017.

[8] Dimitri, Giovanna Maria, and Pietro Lió. "DrugClust: a machine learning approach for drugs side effects prediction." Computational biology and chemistry 68 (2017): 204-210.

[9] Ferdousi, Reza, Reza Safdari, and Yadollah Omidi. "Computational prediction of drug-drug interactions based on drugs functional similarities." Journal of biomedical informatics70 (2017): 54-64.

[10] Harnie, Dries, Mathijs Saey, Alexander E. Vapirev, Jörg Kurt Wegner, Andrey Gedich, Marvin Steijaert, Hugo Ceulemans, Roel Wuyts, and Wolfgang De Meuter. "Scaling machine learning for target prediction in drug discovery using apache spark." Future Generation Computer Systems 67 (2017): 409-417.

[11] Lo, Yu-Chen, Stefano E. Rensi, Wen Torng, and Russ B. Altman. "Machine learning in chemoinformatics and drug discovery." Drug discovery today (2018).

[12] Chen, Ruolan, Xiangrong Liu, Shuting Jin, Jiawei Lin, and Juan Liu. "Machine Learning for Drug-Target Interaction Prediction." Molecules 23, no. 9 (2018): 2208.

[13] Panteleev, Jane, Hua Gao, and Lei Jia. "Recent applications of machine learning in medicinal chemistry." Bioorganic & medicinal chemistry letters (2018).

[14] Kolachalama, Vijaya B., and Priya S. Garg. "Machine learning and medical education." npj Digital Medicine 1, no. 1 (2018): 54.

[15] Lavecchia, Antonio. "Machine-learning approaches in drug discovery: methods and applications." Drug discovery today20, no. 3 (2015): 318-331.

[16] Zhang, Lu, Jianjun Tan, Dan Han, and Hao Zhu. "From machine learning to deep learning: progress in machine intelligence for rational drug discovery." Drug discovery today22, no. 11 (2017): 1680-1685.

[17] Chen, Hongming, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. "The rise of deep learning in drug discovery." Drug discovery today (2018).

[18] Colwell, Lucy J. "Statistical and machine learning approaches to predicting protein-ligand interactions." Current opinion in structural biology 49 (2018): 123-128.

[19] Brown, Nathan, Jean Cambruzzi, Peter J. Cox, Mark Davies, James Dunbar, Dean Plumbley, Matthew A.Sellwood et al. "Big Data in Drug Discovery." In Progress in medicinal chemistry, vol. 57, pp. 277-356. Elsevier, 2018.

[20] Lima, Angélica Nakagawa, Eric Allison Philot, Gustavo Henrique Goulart Trossini, Luis Paulo Barbour Scott, Vinícius Gonçalves Maltarollo, and Kathia Maria Honorio. "Use of machine learning approaches for novel drug discovery." Expert opinion on drug discovery 11, no. 3 (2016): 225-239.