ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

A Review on Healthcare Informatics

Dr. G. T. Prabavathi, M. Shanthipriya

Assistant Professor, Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam, India

Research Scholar, Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam, India

ABSTRACT: There is an enormous increase in electronic health records in healthcare organizations which should be analyzed appropriately to improve patient care which requires high performance computing platforms. Healthcare data contains large quantities of useful information hidden in medical records and big data analytics plays a critical role in predicting disease beforehand. This paper presents an overview of healthcare data types, challenges in healthcare, opportunities and impacts of healthcare analytics. A review on big data analytics research papers that deals with chronic diseases like diabetes, cancer, asthma has been made.

KEYWORDS: Big Data Analytics, Health Care Data types, Healthcare predictive Analytics

I. INTRODUCTION

Big data is the data that exceeds the processing potential of traditional database systems. The data is too big, moves too fast, or doesn't fit the strictures of conventional database architectures (Dumbill 2013). The prominent characteristics of Big data are Volume, Velocity, Variety, Value and Veracity. Big data plays an important role in the field of healthcare.

In healthcare, Big data refers to electronic health data sets that are too large to manage with traditional software or with common data management tools. Big data in healthcare is overwhelming not only because of its volume but also because of the diversity of data types and the speed at which it must be managed. Healthcare analytics finds insights from unstructured, complex and noisy health records of patients for making better health care decisions.

This paper presents an overview of big data in healthcare. Chapter 2 discusses about the various types of healthcare data. The demanding situation of big data in healthcare is discussed in Chapter 3. Chapter 4 provides various impacts of big data analytics in healthcare. Chapter 5 presents a review on some research papers that deals with chronic diseases.

II. HEALTHCARE DATA TYPES

Genomic Data

It refers to genotyping, gene expression and DNA sequence (Chen et al., 2012; Priyanka and Kulennavar, 2014).

Clinical Data and Clinical Notes

Clinical data includes Structured data such as laboratory data (Electronic Medical Records), Unstructured data such as testing reports, patient discharge summaries and Semi-structured data such as data generated by the ongoing conversion of paper records to electronic health and medical records. Literature reveals that 80% of the clinical data are unstructured documents, images and transcribed notes (Yang et al., 2014):

Behavior Data and Patient Sentiment Data

Due to the enormous social media users, web and social media related data in health plan websites also increases. Behavior data includes social media data and streamed data from regular medical monitoring (Ragupathi et al. 2013).

Health Publication and Clinical Reference Data

It includes publications from journals, articles, clinical research/reference materials and text-based practice guidelines.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Administrative, Business and External Data

Healthcare data not only ends just with medical histories of patients but includes business related such as insurance and claims related financial data.

- Biometric data: Fingerprints, handwriting and iris scans, etc.
- Content from medical related e-mails
- Feedback of patients related to treatments

III. CHALLANGES OF BIG DATA IN HEALTHCARE

Large volume, velocity and variety of big data have brought big challenges in data storage, processing, retrieval, search and visualization. Variability and veracity of big data indicate data instability and uncertainty, which often makes Big data analytics difficult. Major challenges of Big Data in medical applications and healthcare are as follows:

1. Unstructured data which includes test results, scanned documents, images and progress notes in the patient's HER are difficult to aggregate and analyze. Efficiently handling large volumes of medical imaging data, extracting potentially useful information and biomarkers and understanding unstructured clinical notes in the right context are challenges (Priyanka and Kulennavar, 2014).

2. From the big data, the focus of the physicians is to find the cause of the disease in order to proceed with effective treatment. Big data means more information, but there is often noisy data or false information (Bottles and Begoli, 2014).

3. In Health Insurance Portability and Accountability Act (HIPAA), the privacy issues are often cited as barriers for collecting big data (Warner, 2013). In tele-cardiology and tele-consultation, data confidentiality in the cloud, data interoperability among hospitals and network latency and accessibility are challenges. (Hsieh et al., 2013)

4. Open access, integration, standardization of readable and useable data is a challenge (White, 2014) due to privacy. Even if there is surety that privacy of patient is protected, Health Care Providers are reluctant to share data due to imbalance between information protection and integrity maintenance

5. The big data have become more damaging by data hackers and data leakage can be costly. There are various instances all over the world where personal data about patients have been hacked by the intruders (Schmitt et al., 2013). Biometric helps to improve the information security and protect against data leakage. However, it is almost impossible to guarantee complete data security.

6. Resource shortfalls such as staffing, budget and infrastructure play a big role as a barrier to the adoption of Big Data by many concerns. Lack of infrastructure, policies, standards and practices were listed as major concerns for big data adaption in healthcare.

IV. OPPORTUNITIES AND IMPACT OF BIG DATA IN HEALTHCARE

The continuing digitization of health records together with the abundant electronic health record (EHR), presents new opportunities to analyze clinical and administrative questions. The opportunities of big data in health care are: (i) Personalized care- Predictive data mining can highlight best practice treatments through early detection and diagnosis and leverage personalized care; (ii) Clinical decision support – Analytics techniques understand, categorize, predict and recommend alternative treatments to clinicians; (iii) Public Health Management – Based on customers search, social content and query, BDA can predict a disease outbreak across patient populations and help to identify the disease trend in a geographical area; (iv) Clinical operations – Mine large amount of historical unstructured data and look for scenarios to predict events. Impacts of big data transformation related to healthcare ecosystem are suggested through the following pathways(Chinchmalatpure, 2016):

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Right Living: Patients can play active role in their own treatment by preventing diseases through appropriate diet and exercise.
Right Care: Patients can get appropriate timely treatment through coordinated approach
Right Provider: Patients should be treated by the right professional properly assigned to the correct match of their disease.
Right Value: Cost effectiveness of health care can be improved by preserving the quality and enhancing financial reimbursement or eliminating fraud and abuse.
Right Innovation: Identification of new therapies and advance as well as boost medical Research and development. Clinical trials and improvement in traditional treatment protocols can be done.

V. RELATED WORK

Umesh et al.,[17] proposed an algorithm implemented in MapReduce for predicting the breast cancer recurrence in SEER dataset. The authors initialized all the unform weight and the weight were updated in every iteration. A base learner function was applied to the weighted form of data, which returns an optimal weak hypothesis, which reduces the weighted error. On each iteration, the authors assigned a weight to the weak classifies. Finally, the algorithm return a final classifier which is a weighted average of all weak classifies. The experiments were demonstrated on Amazon ECZ environment using 17 attributes for predicting classification accuracy. Expectation Maximization algorithm were adopted for an efficient estimation of the unobserved values. The authors demonstrated how to predict whether breast cancer will recur or not using a classification model. The overall cost of the algorithm is O(dn(T+logn)). The main limitation of this algorithm is the training data set to each MapReduce job depends on the recall of the previous MapReduce job. Hence, MapReduce jobs cannot be executed is parallel. Moreover, at each iteration, the data is read from HDFS where the past MapReduce distribution is stored. The overhead for re-stacking and re-processing is more which utilize more CPU resources.

Sivakami[16] proposed a Breast Cancer prediction method using a hybrid technique. The author analyzed the breast cancer data retrieved from Wisconsin dataset where each records in database has an attributes which were found to differ significantly between benign and malignant samples. The author predicted the presence or absence of breast cancer using classifiers by taking the profiles of patients such as age, sex, blood pressure and blood sugar. In this work a hybrid classification which integrates Decision Tree Support Vector Machine algorithm were used. In this methodology classes were selected for the experiments using different age groups, breast size and menopause based on the rules created by ID3 algorithms. First the information were gathered and pre-processed and classified into training and testing data. The author selected 10 parameters for training the feature selection. The performance of the classifiers were measured based on the error rates and accuracy. The parameters for SVM were optimized using Decision Tree. After comparing three classification techniques Instance - Based Learning (IBL), Sequential Minimal Optimization (SMO) and Naïve based classifiers, the experimental results showed that DT-SVM is better than IBL, SMO and Naïve based classifier algorithms.

Asthma is a common disease for both adult and children which requires an accurate prediction of hospitalization of patients. Various reviews have been made to determine the hospitalization decision for asthma patients. Most of the study predict who is likely to suffer from asthma rather than when it is to occur. Joseph et al.,[6] developed an algorithm to predict asthma exacerbation one day in advance based on previous day window using CART. For developing this algorithm, the author used data from home based tele monitoring which was send as daily reports that contained an asthma diary with 22 parameters such as whether the patient has wheeze, cough, cold, chest tightness, shortness of breath, use of inhaler, how many puffs in 24hours/last night, sleeplessness etc. To predict the asthma exacerbation, the authors used four levels: Green zone, High Yellow zone, Low Yellow zone and Red zone to predict whether the patient is doing well, getting worse, dangerous, medical alert. The generalized data set was grouped based on presence or absence of exacerbations at the target day. The CART analyzes were used to construct a binary classification system based on the normal or exacerbation of asthma at target day. At each node, the algorithm selected the variable with the greatest capacity for discriminating between two outcomes such as the presence or absence of exacerbations at the next day. The regression tree branches were added to the tree by the algorithm until a more homogeneous group is reached in terms of the probability of predicting the outcome in question or until few leaves remained. Accuracy, sensitivity and specificity values were used for assessing the CART model in the validation data

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

set by comparing the prediction of future exacerbation with data sets from 7 different scale windows. Resulting algorithm had specificity 0.971, sensitivity of 0.647, and accuracy of 0.809.

J.M. Lillo Castellano et al.,[7] proposed an Arrhythmia classification using Data compression and kernel methods. Arrhythmia database contains cardiac based data. Cardiac method classification (CAC) method has been studied in depth and the author designed a big data analytics method for automatic CAC of EGM (Electrograms) by ICDs (Implantable Cardioverter Defibrillatros) which is based on the effective combination of information theory and kernel methods. They identified that the combined concepts based on data compression with the power of kernel method helps to classify EGMs without large pre-processing methods. The EGMs were extracted from a national scientific big data service using scoop platform for the experiments. The performances of two classifiers are compared in two scenarios using four different input spaces. The results showed KNN can work better than SVM.

Readmission of patients due to improper medication, severity of illness, unmonitored discharge generally occurs in many hospitals. Proper measures are necessary to reduce a risk of readmission. It is essential to develop tools to predict and analyze the readmitted patient records. Diabetes is a chronic disease and plays a higher role of patient readmission in many hospitals. Saumya salian et al.,[15] reviewed a effectiveness of big data in predicting the risk of readmission of diabetic patients. The data collected from UCI repository where preprocessed for training predictive models using 32HDFS in Hadoop framework. Class labels where identified using two variables namely *0* which indicates tested negative for readmission and *1* which indicates tested positive for readmission. The classification models Logistic Regression, Decision Tree, SVM where applied to perform the experiment and proved that DT and SVM give misclassification error rates. Confusion matrix and Formulation matrix were used to represent predicted classification and features relevant of readmission labels. The authors identified the chance of readmission of diabetic patients can be successfully identified using classification rules with plasma glucose as an important parameter.

There is an enormous increase of social media users to interact with each other regarding health information. Recent research shows a growing trend of sharing health related information in social media such as Facebook, Twitter., etc. The socially shared health information seeks to understand our health situation. Hence scientists have shown interest in investigating the importance of social life and health situation. Nadiya et al.,[10] performed data analytics on public health from Facebook engagement, The authors applied machine learning techniques on big social data of public health Facebook to understand the users posts. Their research focused on studying impact of healthcare posts on Facebook. The data was collected from Facebook walls from various public health organizations such as international agencies, individual bloggers using social data analytics tool which included 16 direct and indirect attributes and 9 derived attributes . First the correlation between attributes Page Likes, Talking About, Post Share, Post Likes, Comments, Comments Likes, Comment Reply and Comment Reply like were identified. Four attributes Post Share, Post Like, Comment and Comment Like which were most correlated were chosen to represent Post engagement. Using K-Means clustering, a clear picture of popular and less popular post are identified. The results of clustering and statistical analysis showed that in order to achieve better engagement, organization should represent their posts through photos instead of common information sharing techniques. The authors suggested that organizations must avoid status update and think about better visual representation.

Due to climatic change nuclear power generation, there arises possibilities of large scale epidemic and there is an increasing tendency of monitoring epidemic trend from the big data through internet. One common analysis is monitoring the influenza trend from news, blogs and data from internet. Hideo et al.,[4] analyzed English twitter data to extract the keywords related to influenza. The authors analyzed the possibility of building a multiple linear regression model with ridge regularization by combining twitter messages and identified the CDC's ILI data outperforms a single linear regression model.

Various disease prediction big data analytics researches select the characteristics automatically from large number of data rather than previously selected characteristics and also consider only structured data. Min Chen et al.,[9] combined the structured and unstructured medical records collected from central china hospital. The laboratory data and patients information were treated as structured data and doctors interrogation of patients records and diagnosis were considered as unstructured data. The latent factor model was used to reconstruct the missing data to overcome the difficulty of incomplete data. To extract meaningful features from structured data, consultation with hospitals experts were made and to predict the features from unstructured data, the authors used Conventional Neural Network (CNN) algorithm. Using the Novel CNN based multi model risk disease risk prediction (CNN-MDPR) algorithm, chronic disease outbreaks in disease frequent communities were predicted. The goal of their research was to predict whether a patient is

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

at the risk of disease or not. The algorithm resulted with 94.8% accuracy which is faster than CNN-based Unimodal disease risk prediction (CNN-UDRP) algorithm.

VI. CONCLUSION

Due to the rapid growth of healthcare data, healthcare providers and medical practitioners should handle, extract knowledge form the healthcare records to enhance advance disease prediction. An overview of big data and its types is outlined followed by the challenges of healthcare analytics. Few research papers that deal with diseases such as diabetes and cancer have been reviewed.

REFERENCES

1. Bottles, K. and E. Begoli:"Understanding the pros and cons of big data analytics", 2014. Physician Exec., 40: 6-12.
2. Chen, H.C., R.H.L. Chiang and V.C. Storey: "Business intelligence and analytics: From big data to big impact" , 2012.. MIS Q., 36: 1165-1188.
3. Dumbill, E., 2013. Making sense of Big Data.
4. Hideo Hirose, Liangliang Wang : " Prediction of Infectious Disease Spread using Twitter : A Case of Influenza ", 2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming.
5. Hsieh, J.C., A.H. Li and C.C. Yang, 2013. Mobile, cloud and big data computing: Contributions, challenges and new directions in telecardiology. Int. J. Environ. Res. Public Health, 10: 6131-6153. DOI: 10.3390/ijerph10116131
6. Joseph Finkelstein and In Cheol Jeong : " Using CART for Advanced Prediction of Asthma Attacks Based on Telemonitoring Data", 978-1-5090-1496-5/16, 2016 IEEE
7. Lillo-CastellanI. Mora-Jimenezo J. M, I. Moreno-Gonzalez, M. Montserrat-Garcia-de-Pablo, a. Garcia-Alberola and J. L. Rojo-Alvarez: " Big-Data Analytics for Arrhythmia Classification using Data Compression and Kernel Methods", ISSN : 2325-8861, Computing in Cardiology 2015; 42:661-664
8. Madhura A. Chinchmalatpure, Dr. Mahendra P. Dhore: " Review of Big Data Challenges in Healthcare Application ", IOSR Journal of computer Engineering 2016, e-ISSN: 2278-8727, e-ISSN : 2278-8727, PP. 06-09
9. Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang: " Disease Prediction by Machine Learning Over Big Data From Healthcare Communities", Special Section on Healthcare Big Data, IEEE Access, Volume 5, April 2017.
10. Nadiya Straton, Kjeld Hansen, Raghava Rao Mukkamala, Abid Hussain, Tor-Morten Gronli, Henning Langberg and Ravi Vatrabu: " Big Social Data Analytics for Public Health : Facebook Engagement and Performance ", IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom), 2017
11. Priyanka K, Nagarathna Kulennavar : "A Survey on Big Data Analytics In Health Care ", International Journal of Computer Science and Information Technologies, Vol. 5(4), 2014, 5865-5868
12. Ragupathi W and V. Ragupathi : " An Overview of Health Analytics ", Vol. 4 Issue 3, ISSN : 2157-7420
13. Rishika Reddy A and P. Suresh Kumar: "Predictive Big Data Analytics in Healthcare ", 2016 Second International Conference on Computational Intelligence & Communication Technology.
14. Schmitt, C., M. Shoffner and P. Owen:" Security and privacy in the era of big data: The SMW, a technological solution to the challenge of data leakage", 2013. RENCI, University of North Carolina at Chapel Hill.
15. Saumya Salian, Dr. G. Harisekaran : " Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients", International Journal of Science and Research(IJSR), Vol. 4, Issue. 4, April 2015
16. Sivakami K : "Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model", International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-5, August 2015, ISSN: 2395-3470
17. Umesh D. R and B. Ramachandra : " Big Data Analytics to Predict Breast Cancer Recurrence on SEER Dataset using MapReduce Approach", International Journal of Computer Applications(0975-8887), Volume 150-No. 7, September2016.
18. Warner, D : "Safe de-identification of big data iscritical to healthcare". Health Information Management, 2013.
19. White, S.E. : "A review of big data in health care: Challenges and opportunities". Open Access Bioinform., 6: 13-18. DOI: 10.2147/OAB.S50519, 2014
20. Yang, S., M. Njoku and C.F. Mackenzie: "Bigdata approaches to trauma outcome prediction and autonomous resuscitation", 2014.