Disease Prediction Using Big Data and Machine Learning

Prakash Chand Sharma[1], Surbhi Singh[2], Archna Goyal[3], Er. Saurabh Sharma[4]
[1,2,3] *Asst.Professor, Aryabhatta College of Engineering & Research Center, Ajmer*
[4]*HOD, Aryabhatta College of Engineering & Research Center, Ajmer*

***Abstract*- Now a days big data is the fastest and more widely used in every field .With the help of big data medical and health care sectors achieves their growth and with help of big data benefit of a accurate medical data analysis , early disease prediction, accurate data of an patient can be securely stored and used .Moreover the accuracy of an analysis can be reduced due to an various reason like incomplete medical data, some regional disease characteristics which can be outbreaks the prediction. In this paper we can use a machine learning algorithm for the accurate disease prediction for that purpose we can collect the hospital data of a particular region. For missing data we can use latent factor model to achieve the incomplete data. In the previ-ous work for disease prediction Convolutional Neural Network Based Unimodel Disease Prediction (CNN-UDRP) Algorithm is used. Convolutional Neural Network Based Multimodal Disease Prediction(CNN-MDRP) algorithm is overcome the drawbacks of CNN-UDRP algorithm only focus work on a structured data but CNN-MDRP algorithm uses both structured and unstructured data from the hospital. None of the existing work focused on both data types in the area of medical big data analysis .CNN-MDRP algorithm prediction is more accurate than compared to the previous prediction algorithm.**

***Index terms*- Data analytics, Health care data, Machine learning.**

I.INTRODUCTION

The concept of the big data is not a new concept it is constantly changing. Big data is nothing but the collection of data. There are three important vs of data that is velocity, volume and variety. Healthcare is a best example of three v/s of data. The healthcare data is spread among the multiple medical systems, healthcare sectors, and government hospitals with the benefits of a big data more attention is paid to the Disease Prediction. Number of researches has been conducted to selecting the characteristics of a disease prediction from a large volume of a data. Most of the existing work is based on a structured data. For the unstructured data one can use a convolutional neural network. Convolutional neural network are made up of a neurons, each neurons receives some inputs and performs operations and the whole network expresses a single differentiable score functions. The accuracy of a disease prediction can be reduced because
there is a more difference in a various regional disease because of climate and living habits of the peoples in their particular regions .However there are more challenges remain that are: 1) How should missing data is collected? 2) How should certain regional characteristic of disease can be deter-mined? 3) How should overcome the climate and living habits problems? To reduce this challenges we combine both the structured and unstructured data.to accurately predict the disease overcome the problem of a missing and incomplete data .we can use a latent factor model. In the previous work only structured data can be used but for the accurate results we can use the unstructured data. We can select characteristic automatically using CNN algorithm. WE can purpose a CNN-MDRP algorithm for both the data types. We can use machine learning algorithm for more accurate results.

OBJECTIVES 1. Disease Prediction for Structured Data we use traditional machine learning algorithms i.e., Naïve Bayesian (NB).

II. REVIEW OF LITERATURE

In this paper, the authors propose a convolutional neural network based multimodal disease risk prediction (CNNMDRP) .This algorithm overcome the drawbacks of (CNN-UDRP) convolutional neural network based unimodal disease risk prediction. This algorithm uses both the structured and unstructured data of a hospital. None of the existing algorithm can work on both the structured and unstructured data. Its accuracy is about 94.8 In this paper, the researchers

present how artificial intelligence applied to medical field for the efficient diagnosis.

For that purpose they use a k nearest neighbours algorithm and they check the accuracy of the algorithm with the help of UCI machine learning repository datasets. They had to generate patients input and test data for diagnosis. They use a real patient data. They add a additional training sets allow more medical conditions to be classified with the minimal no of changes to the algorithm. [2]

In this paper, they applying a machine learning techniques by using EMC'S from outpatients department and the algo-rithm are based on a DNN AND DBDT, It can be achieve a high UAR for predicting the future stroke prediction. It provides a several advantages like high accuracy, fastest prediction, and consistency of results. DNN algorithm also requires a lesser amount of data. DNN algorithm can achieves a optimal results by using a lesser amount of a patient data than compared to the GDBT algorithm. [3]

In this paper, distributed computing environment processing the large volume of a data is done based on Map Reduce. To find the accuracy of a patient data the classification is used. In this paper more focused on find out the nearest accuracy of a classifiers. The CART model and random forest is built for the data and accuracy of the classifier

is found. By using the random forest algorithm they can found the more nearest accuracy of the prediction. The prediction analysis helps to the doctors to identify the patient's admissions on to the hospital. Predictive model using scalable random forest classification which can accurately give the result rate of risk. [4] In this paper for heart disease prediction they use a Neavi Bayes and Decision tree algorithm. They used a PCA to reduce the no of attributes, after reducing the size of the datasets; SVM can outperform a Neavi Bayes and Decision tree. SVM can also be used for prediction of hearts disease. The main goal of this paper is to predict the diabetics disease. Using a WEKA data mining tools. Data mining is very useful techniques used by health care sector for classification of disease. The aim of this paper is to study supervised machine learning algorithm to predict the heart disease. [5] In this paper the data mining and the big data in the healthcare sector is introduced. Machine learning algorithm has been used to study the healthcare data. The continuous increase of data in a healthcare. Several countries are spending a lot of resources, scientist leads to fix the problem of storage and organization of data the data mining will help

exploitation complexity of the data and find out the new result this paper is based on the use of data mining and big data in the healthcare sector. [6] Traditional wearable devices have various shortcomings, such as comfortableness for long-term wearing, and insufficient accuracy, etc. Thus, health monitoring through traditional wearable devices is hard to be sustainable. In order to obtain healthcare big data by sustainable health mon-itoring, we design Smart Clothing, facilitating unobtrusive collection of various physiological indicators of human body. To provide pervasive intelligence for smart clothing system, mobile healthcare cloud platform is constructed by the use of mobile internet, cloud computing and big data analytics. This paper introduces design details, key technologies and practical implementation methods of smart

clothing system. Typical applications powered by smart clothing and big data clouds are presented, such as medical emergency response, emotion care, disease diagnosis, and real-time tactile interaction. Especially, electrocardiograph signals collected by smart clothing are used for mood monitoring and emotion detection. Finally, we highlight some of the design challenges and open issues that still need to be addressed to make smart clothing ubiquitous for a wide range of applications[7]

III. SYSTEM ARCHITECTURE / SYSTEM OVERVIEW

A. Proposed work

In a proposed system we can first get the large volume of a healthcare big data, then that data is considered as training data. Naive Bayes algorithm is used for the clarification of the data. Then after the clarification the hospital data similar type of data can be stored. Then CNN extract the text characteristics automatically. In that we use a CNN MDRP algorithm that uses both structured unstructured hospital data. Selecting the characteristics automatically form a large number of data. This improves the disease prediction rather than previously selected characteristics. CNN- MDRP

algorithm helps to accuracy of the result of a disease prediction over a large volume of data from hospital

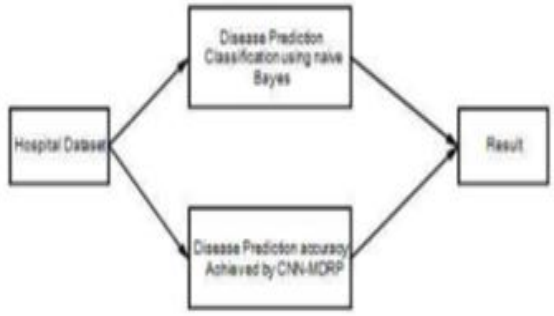

Fig. 1. Block Diagram of Proposed System (Proposed System Architceture)

B. ALGORITHM Used machine learning algorithm: - Naive Bayes

It is a classification technique based on a Bayes theorem. Javier Bayes algorithm is easy to build and mainly useful for a very large amount of data sets. In a naive Bayes it can convert the data set in a frequency table and then create a likelihood table by finding the probabilities like overcast probability. In our paper we are using the naive Bayes algorithm for the accurate outcome of prediction from the large volume of a medical data. Bayes theorem provides a way of calculating the posterior probability, P(c—x), from P(c), P(x), and P(x—c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence P(c—x) is the posterior probability of class (target) given predictor (attribute). P(c) is the prior probability of class. P(x—c) is the likelihood which is the probability of predictor given class. P(x) is the prior probability of predictor Where C and X are two events (e.g. the probability that the train will arrive on time given that the weather is rainy). Such Nave Bayes classifiers use the probability theory to find the most likely classification of an unseen (unclassified) instance. The algorithm performs positively with categorical data but poorly if we have numerical data in the training set.

IV. SYSTEM ANALYSIS

A. Data set:

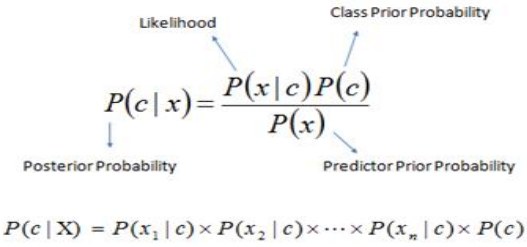

Fig. 2. (Equation Of Naive Bayes)

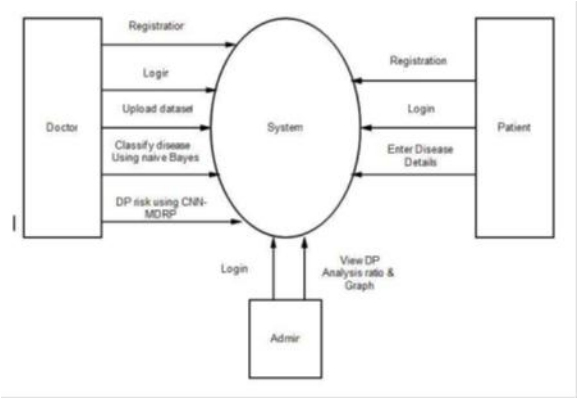

Fig. 3. Data Flow Daigram (Data flow Daigram of system)

a. Hospital data A large volume of datasets of a patient can be given by a hospital and the data can be stored in the data centre to protect the patient privacy and security of stored data, we create a security access mechanism.
b. Structured data The structured data is nothing but the laboratory data, patients basic information like patients age, gender, life habits, height, weight etc.
c. Unstructured Data Unstructured Data is a data of patients medical history, patients illness, and doctors interrogation and diagnosis. The 20 hospitals datasets consisting 20,000 documents and data of patients. The 20 hospital dataset is a popular dataset for experiments in application of a machine learning techniques

B. Required configuration:

For implementation the proposed system required a follow-ing configuration is required

1. JDK 1.8 2. Database - Mongo DB 3. Server-Apache Tomcat server

C. Data Imputations:

There is a large number of missing data due to the human error. This we need to fill the structured data.

Before data Imputations for that we have to first identify uncertain or incomplete medical data and then modify or delete. Then to improve the quality of data and then we use data integration for data processing.

V. RESULT

Expected result in this section table 1 represent difference between the existing system result and proposed system result proposed system can predict disease fast and more accurate than the existing system. for disease risk modelling, the accuracy of risk prediction depends on the diversity feature of the hospital data, i.e., the better is the feature description of the disease, the higher the accuracy will be. For some simple disease, only a few features of structured data can get a good description of the disease, resulting in fairly good effect of disease risk prediction. But for a complex disease, such as cerebral infarction mentioned in the paper, only using features of structured data is not a good way to describe the disease.

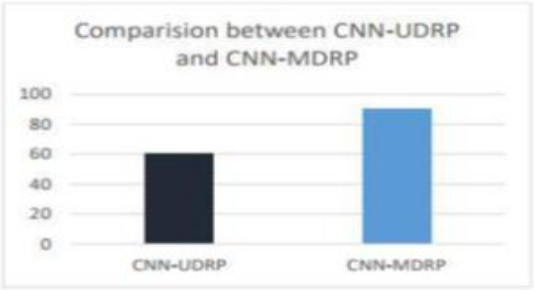

Fig. 4. Comparative Result (comparison of CNN-UDRP AND CNN-MDRP)

Sr.No	Existing Method Result	Proposed Method Result
1	60	90

Fig. 5. Comparative Result (Comparative Result)

VI. CONCLUSION

In this paper we propose a CNN-MDRP algorithm for a disease prediction from a large volume of hospital's structured and unstructured data. Using a machine learning algorithm (Neavi- Bayes) Existing algorithm CNN-UDRP only uses a structured data but in CNN-MDRP focus on both structured and unstructured data the accuracy of disease prediction is more and fast as compared to the CNN-UDRP. By combining the structured and unstructured data the accuracy rate can be reach to 94.80

REFERENCES

[1] in Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang, Disease Prediction by Machine Learning over Big Data from Healthcare Communities, 2169-3536 (c) 2016 IEEE.

[2] hahab Tayeb*, Matin Pirouz*, Johann Sun1, Kaylee Hall1, Andrew Chang1, Jessica Li1, Connor Song1, Apoorva Chauhan2, Michael Ferra3, Theresa Sager3, Justin Zhan*, Shahram Latifi, Toward Predicting Med-ical Conditions Using k-Nearest Neighbours, 2017 IEEE International Conference on Big Data.

[3] hen-Ying Hung, Wei-Chen Chen, Po-Tsun Lai, Ching-Heng Lin, and Chi-Chun Lee, Comparing Deep Neural Network and Other Machine Learning Algorithms for Stroke Prediction in a Large-Scale Population-Based Electronic Medical Claims Database, 2017 IEEE.

[4] reekanth Rallapalli Faculty of computing Botho University Gaborone, Botswana, Predicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm, 2016 IEEE.

[5] rof. Dhomse Kanchan B. Assistant Professor of IT department METS BKC IOE, Nasik Nasik, Mr. Mahale Kishor M. Technical Assistant of IT department METS BKC IOE, Nasik, India, Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analy-sis,2016 IEEE.

[6] oubida Alaoui Mdaghri, Mourad El Yadari, Abdelillah Benyoussef, Ab-dellah El Kenz Faculty of Science Rabat Morocco, Rabat, Study and analysis of Data Mining for Healthcare, 2016 IEEE.

[7] Chen, Y. Ma, J. Song, C. Lai, B. Hu, Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring, ACM/Springer Mobile Networks and Applications Vol. 21, No. 5, pp.825C845, 2016

[8] Richard Osuala and Ognjen Arandjelovi c University of St Andrews, United Kingdom, Visualization of Patient Specic Disease Risk Prediction, 2017 IEEE.