# Disease Prognosis by Machine Learning Over Big Data from Healthcare Communities

**M.Rajeswari, A.Chandrasekar, Nasiya PM**

***Abstract:*** ***With huge information headway in biomedical and healthcare communities, appropriate examination of therapeutic information helps early sickness identification, tolerant consideration and network administrations. Prediction accuracy is diminished when the nature of medicinal information is inadequate. At that point the various areas appear, one of kind qualities of certain local infections, which may debilitate the expectation of illness episodes. In this paper, machine learning method is applied for viable forecast of interminable disease in the history of predicting diseases. The main intension is to have different prediction models over genuine medical clinic information gathered from focal China in 2013-2015. To conquer the trouble of deficient information, a latent factor model is used to regenerate the irrecoverable data. Here, experiment on a territorial chronic infection of cerebral localized necrosis is done. CNN-MDRP (convolutional neural system based multimodal infection chance prediction) algorithm is explained utilizing organized and unstructured information from medical clinic. Apparently, none of the current work establishes on the two information types in the zone of therapeutic enormous information investigation. Contrasted with numerous prediction algorithm, the precision accuracy of the proposed method arrives at 94.8% with a combination speed which is quicker than that of the CNN-UDRP(convolutional neural network based unimodal disease risk prediction) technique.***

***Keywords: Big data analytics; Machine Learning; Healthcare***

## I. INTRODUCTION

Concept of the big data is not a new concept which is constantly changing. Big data is nothing but the collection of data. There are three important v's of data such as velocity, volume and variety. Healthcare is a best example of adapting these 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 in which more attention is paid to the Disease Prediction. Numbers of researches have been conducted to selecting the characteristics of a disease prediction from a large volume of a data. Most of the existing works were based on a structured data. For the unstructured data, one can use a convolutional neural network which is 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.

## II. RELATED WORKS

A. In [1], the authors proposed a CNNMDRP (convolutional neural network based multimodal disease risk prediction) which overcomes 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 compared with other existing algorithm which can work on either the structured or unstructured data. Authors have explained that the proposed algorithm produced the accuracy of 94.8. In this paper, the researchers presented how artificial intelligence applied to medical field for the efficient diagnosis. Also, to fulfill this need, authors used a k nearest neighbour's algorithm and verified the accuracy of the algorithm with the help of UCI machine learning repository datasets. In addition to this,it is needed to generate patients input along with test data for diagnosis. Authors have considered a real patient data for which additional training sets were added which allow more medical conditions to be classified with the minimal no of changes in the algorithm. In this paper [2], authors have applied a machine learning techniques by using EMC'S from outpatients department and the algorithm were based on a DNN AND DBDT which can achieve a high UAR for predicting the future stroke problem. This technique provides a several advantages such as high accuracy, fastest prediction, and consistency of results.

DNN algorithm also requires a lesser amount of data that can achieve a higher impact while applying a insignificant amount of a patient data compared to the GDBT algorithm. In this paper [3], distributed computing environment has been detailed which processes the large volume of a data based on Map Reduce. The CART model together with random forest was constructed for the information and exactness of the classifier was established.

**Revised Manuscript Received on November 15, 2019.**

**Dr.M.Rajeswari***, Professor of Department of CSE, Sahrdaya College of Engineering and Technology. Thrissur, Kerala, India. Email: rajeswarim@sahrdaya.ac.in

**Dr. A.Chandrasekar**, Professor of Department of CSE Malla Reddy Institute of Technology and Science, Secunderabad. Email: chandru.as76@gmail.com

**Nasiya PM**, Assistant Professor of BCA, MES Asmabi College, P.Vemballur, Kerala, India. Email: nasiya.pm2010@gmail.com

By utilizing the arbitrary timberland calculation they can found the closest exactness of the forecast. The prediction analysis helps to the doctors to identify the patient's admissions on to the hospital. It is noted that prescient model utilizing versatile random forest order which can precisely produce the outcome of hazard. In this paper [4], authors applied a Naive Bayes and Decision tree algorithm for heart disease prediction. They used a PCA to minimize the no of attributes upon reducing the size of the datasets; SVM can outperform a Naïve Bayes and Decision tree. Authors elaborated that SVM can also be utilized for prediction of hearts disease. The main goal of this paper is to determine the diabetics disease using a data mining tool named as WEKA. Data mining is a very useful technique which is applied regularly by health care sector for the classification of and determination of disease. The aim of this work is to adapt supervised machine learning algorithm to forecast the heart disease. In [5], the data mining along with big data in the healthcare sector was elaborated for which Machine learning algorithm has been used to examine the healthcare data. The continuous increase of data in a healthcare sector, several countries is spending a lot of resources, scientist help to cure the issues of space and establishment of data. Also, data mining will help exploitation complexity of the data and find out the new result which is based on the use of data mining and big data in the healthcare sector. Conventional wearable devices [6] have different deficiencies, for example, ease for long haul wearing, and deficient exactness, and so on.

## III. PROPOSED METHOD

### A. CNN-based Multimodal Disease Risk Prediction (CNNMDRP) Algorithm

CNN-UDRP utilized uniquely for the content information to foresee whether the patient is at high danger of cerebral dead tissue. Concerning organized and unorganized content information, a CNNMDRP method dependent on CNN-UDRP has been proposed .The handling of content information is comparative with CNN-UDRP which can remove 100 highlights about content informational index. For structure information, we extricate 79 highlights. At that point, we direct the component level combination by utilizing 79 highlights in the S information and 100 highlights in T-information. For computation methods, full connection layer are similar with CNNUDRP algorithm. Figure 1 shows the disease prediction model using various classification algorithms. In CNN-MDRP algorithm, there are two divisions of the training process which is elaborated below,

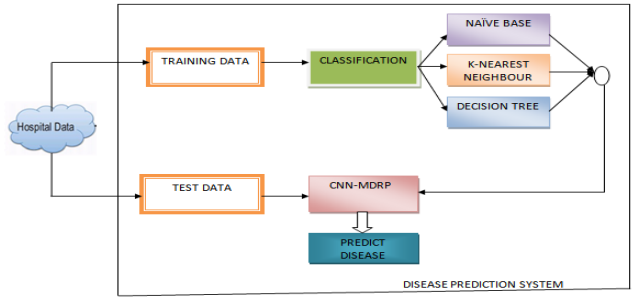


**Fig.1Disease prediction model**

*Training word Embedding:*

Word vector preparing requires unadulterated corpus, that is, it is smarter to utilize an expert corpus. In this paper, we separated the content information of all patients in the clinic from the therapeutic huge server farm. Subsequent to cleaning the data, it is desirable to assign them as corpus set. Utilizing ICTACLAS word division device, word2vector device n-skip gram calculation prepares the word vector, word vector measurement is set to 50. Experiment results show that around 52100 words in the word vector has been finalized after the training.

*Training parameters of CNN-MDRP:*

Stochastic inclination strategy is utilized to prepare parameters to arrive at the hazard appraisal whether the patient experiences cerebral localized necrosis. Some propelled highlights will be tried in future investigation, for example, fractal measurement, orthogonal wavelet change and so forth. The algorithm act as positively with categorical data but poorly if numerical data in the training set.

a) Hospital data: A large volume of datasets of a patient can be given by a hospital which can be processed in the information centre to preserve the patient privacy and authentication of stored data, a security access technique has been created.

b) Structured data: The structured data is nothing but the laboratory data, patient's basic information like patients age, gender, life habits, height, weight etc.

c) Unstructured Data: Unstructured Data is a data of patient's medical history, patient's illness, and doctor's 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

The proposed system requires a following configuration for implementation

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

### C. Data Imputations

There is an enormous number of missing information because of the human blunder which requires the information to be processed in an organized manner. Also, it is required to initially distinguish unsure or inadequate therapeutic information before performing information imputation which further would be modified. Also, to improve the nature of information reconciliation is required for information preparing.

## IV. ANALYSIS OF OVERALL RESULTS

### A. Structured Data (S-data)

For S-information, customary AI calculations are being applied, i.e., NB, KNN and DT calculation to foresee the danger of cerebral localized necrosis infection. NB characterization is a basic probabilistic classifier which requires to figure the likelihood of highlight properties. The KNN algorithm is utilized for a preparation of data set, and the nearest k case in the preparation

informational collection is found. It is desirable for KNN to decide the separation and determination of k esteem. The information is standardized from the outset to utilize the Euclidean separation to quantify the distance. To decide the best classifier and the exactness of the model, the 10-overlap cross-validation technique is utilized for the preparation set, and information from the test set are not utilized in the preparation time. The model's fundamental structure is appeared in Fig. 2. It is desirable to notice that the exactness of the three AI calculations .The precision of DT which is 63% is most noteworthy, trailed by NB and KNN.

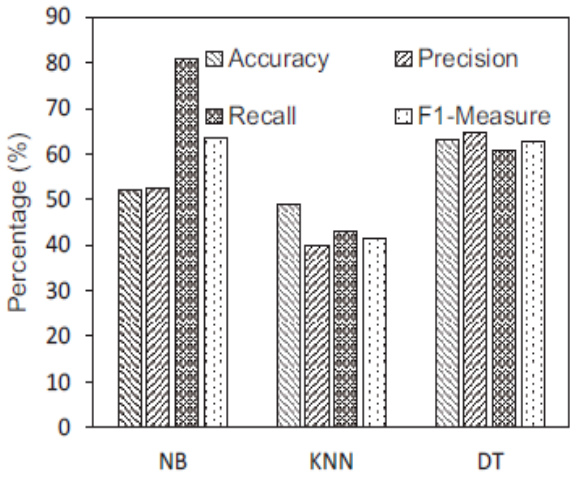


**Fig .2 Comparisons of accuracy, precision, recall and F1-Measure under S-data for NB, KNN and DT**

### B. Structured and Text Data (S&T-data)

In this experiment, the text feature is 100 and selected number of words is 7. For CNN-UDRP (T-data) and CNNMDRP (S&T-data) algorithms, the effectiveness of CNN-UDRP (T-data) and CNN-MDRP (S&T-data) techniques produce little variance in terms of results.

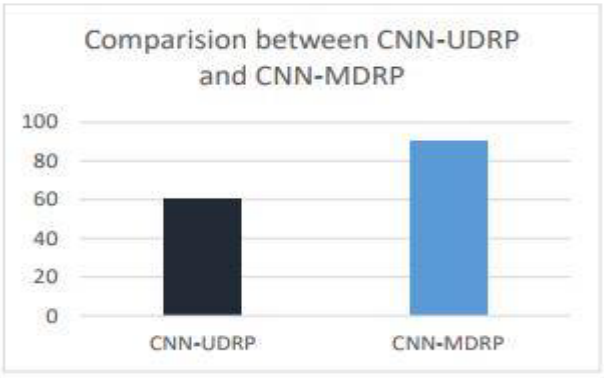


**Fig .3 Comparisons of CNN-UDRP AND CNN-MDRP**

**Table- I: Comparative results**

| | CNN-UDRP | CNN-MDRP |
|---|---|---|
| **Comparison Results** | 60 | 94.80 |

In summary, the performance of CNN-MDRP (S&T-data) is better than CNNUDRP (T-data).prediction shown in figure 3. Yet, for an intricate illness, for example, cerebral localized necrosis referenced in the paper, just utilizing highlights of organized information is anything but a decent method to depict the malady. Table 1 describes the comparative results of the same.

## V. CONCLUSION

In this paper, a CNN-MDRP methodology has been applied for a disease forecast from a huge volume of emergency clinic's organized and unstructured information. Machine learning technique such as Naive-Bayes was utilized for existing calculation CNN-UDRP just uses an organized information however in CNN-MDRP concentrated on both organized and unstructured information for which the exactness of disease prediction is good and quick when contrasted with the CNN-UDRP. By combining the structured and unstructured data the accuracy rate has reached to 94.80.

## REFERENCES

1. Min 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 rallapalli.sreekanth@bothouniversity.ac.bw, Predicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm, 2016 IEEE.
5. Prof. Dhomse Kanchan B. Assistant Professor of IT department METS BKC IOE, Nasik Nasik, India kdhomse@gmail.com , Mr. Mahale KishorM. Technical Assistant of IT department METS BKC IOE, Nasik, India kishu2006.kishor@gmail.com, Study of Machine Learning Algorithms forSpecial Disease Prediction using Principal of Component Analy-sis,2016 IEEE.
6. 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,
9. United Kingdom, Visualization of Patient Specic Disease Risk Prediction,2017 IEEE.

## AUTHORS PROFILE

**Dr. M. Rajeswari** is currently working as an Associate Professor in the Department of Computer Science and Engineering in Sahrdaya College of Engineering and Technology, Thrissur, Kerala. She has received Ph.D degree in Information and Communication Engineering from Anna University, Chennai. She is having more than 12 years of experience in teaching. She has published more than 20 papers in various International Journals and

presented more than 14 papers in both national and International Conferences. She has written a chapter in a book named as Recent Development in Wireless Sensor and Ad-hoc Networks by Springer Publication. She is acting as a Reviewer of Computational Intelligence and Neuroscience journal by Hindawi Publication and International Journal of Wireless Personal Communication by Springer Publication. She has given a Guest Lecturer in various subjects such as Computer Architecture, System Software, Theory of Computation, Formal Languages and Automata Theory and Operating Systems in Premier Institutions. She has also given an invited talk in Anna University Sponsored FDP on Programming and Data Structures – II. She has acted as a jury in a National level project fair, Symposium and Various International Conferences. She has been awarded as a best faculty in the year 2010-2011 in Angel College of Engineering and Technology, Tirupur, Tamilnadu. She has attended various Seminars, Workshops and Faculty Development Programmes to enhance the knowledge of student's community. She is also an active life time member in Indian Society of Technical Education.

**Dr.A.Chandrasekar** received B.Sc. Degree in Computer Science from Nagamalai Navarasam Arts and Science College, Bharathiar University, Tamil Nadu, India in 1998, M.Sc. Degree in Computer Technology from K.S.R. College of Technology, Anna University, Tamil Nadu, India in 2000, M.E. in Computer Science and Engineering from K.S.R. College of Technology, Anna University, Tamil Nadu, India in 2006. He also obtained his Ph.D. Degree in Information and Communication Engineering from Anna University, Tamil Nadu, India in 2016. He is having 16 years of teaching experience in various institutions. He has published 17 papers in various International Journals. His area of interest includes Mobile Computing, Design and Analysis of Algorithms and Internet of Things.

**Nasiya PM** has completed her M.Tech – CSE in Sahrdaya College of Engineering and Technology, Thrissur, Kerala. She is working as an Assistant Professor in the department of BCA at MES Asmabi College, P.Vemballur.