International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

PREDICTION OF PROBABILITY OF DISEASE BASED ON SYMPTOMS USING MACHINE LEARNING ALGORITHM

Harini D K[1], Natesh M[2]

[1]*M.Tech Student, Dept. of Computer Science & Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India*
[2] *Associate Professor, Dept. of Computer Science & Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India*

Abstract - Big data has a major impact on healthcare analytics and it have the capacity to reduce costs of treatment, predict outbreaks of epidemics, avoid preventable diseases and improve the quality of life. Accurate analysis of medical data benefits in early disease detection and well patient care in big data. The analysis accuracy is reduced when we have incomplete data. In this paper, machine learning algorithms is used for effective prediction of diseases. Latent factor model is used to overcome the difficulty of missing data. A new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm is proposed in this paper. It uses both structured and unstructured data from hospital for effective prediction of diseases.

***Key Words*: Big data, Machine learning, Disease Prediction, CNN-MDRP, Structured data, Unstructured data.**

1. INTRODUCTION

The 21st century Industries are generating more data that will grow faster. The organizations utilize this data to make important decisions. The industries that generate huge data are Hospitals, Educational Institutions and many other companies. Healthcare is one among the top that generates large amount of data. Here we apply Machine learning algorithms to maintain complete hospital data. Machine learning allows building models to quickly analyze data and deliver results, leveraging both past and real-time data. With machine learning techniques, doctors can make better decisions on patient's diagnoses and treatment options, which leads to the improvement of healthcare services.

Healthcare is a prime example of how the three dimensions of Big data is used. First is velocity, second is variety and third one is volume. These types of data is spread among multiple healthcare systems, health insurances, researchers, government entities and so forth.

Various researches have been conducted to improve the accuracy of risk classification from a large data. The existing work will consider only structured data. For unstructured data, convolutional neural network (CNN) is used to extract text characteristics automatically. But none of previous work handles medical text data by CNN and also there is a large difference between diseases in different regions, because of the diverse climate and living habits in the region [1].

2. LITERATURE SURVEY

Beforehand, medical team was trying for human services experts to gather and examine the enormous volume of information for powerful expectations and medicines. Since around then there were no advancements or apparatuses are accessible for them. Presently, with machine learning, we make it moderately simple. Huge information advancements, for example, Hadoop are all the sufficiently more for wide-scale selection. Indeed, 54% of associations are utilizing Hadoop as large information handling instrument to get data in human services. 94% of Hadoop clients perform investigation on voluminous information. Machine learning calculations can likewise be useful in giving essential insights, constant information and progressed examination as far as the patient's malady, lab test comes about, circulatory strain, family history, clinical trial information and more to specialists.

Human services framework creates expansive measure of information, the test is to gather this information and successfully utilize it for investigation, forecast and treatment. The principle way to deal with human services framework is to keep the sickness with early location instead of go for a treatment after conclusion. Customarily, specialists utilize a hazard number cruncher to survey the likelihood of sickness advancement. These adding machines utilize central data like socioeconomics, medicinal conditions, life schedules and more to figure the likelihood of advancement of a specific malady. Such counts are finished utilizing condition based scientific techniques and devices. The issue with this technique is the low exactness rate with a comparable condition based approach. Be that as it may, with late improvement in advancements, for example, huge information and machine taking in, it's conceivable to get more precise outcomes for illness expectation.

Doctors are collaborating with analysts and PC researchers to grow better instruments to foresee the sicknesses. Specialists in this field are chipping away at the procedures to recognize, create, and tweak machine learning calculations and models that can convey exact forecasts. To build up a solid and more precise machine learning model, we can utilize information gathered from considers, quiet

socioeconomics, restorative wellbeing records, and different sources.

The distinction between conventional approach and the machine learning approach for sickness forecast is the quantity of ward factors to consider. In a customary approach, not very many factors are viewed as, for example, age, weight, stature, sex, and then some. Then again, machine learning can think about countless, which brings about a superior precision of social insurance information. As indicated by a current report, the analyst got better symptomatic precision, utilizing whole medicinal records by considering around 200 factors.

Aside from illness expectation, there are the couple of more potential zones like medication revelation or electronic wellbeing records where machine learning can enhance social insurance industry. With machine learning applications, the human services and solution fragment can progress into another domain and totally change social insurance tasks.

Subsequently the grouping of danger of infections in light of enormous information investigation has the accompanying difficulties: How should the missing information be tended to? In what capacity should the fundamental incessant maladies in a specific locale decided? By what means can huge information examination innovation be utilized to break down the infection and make a superior model? To tackle these issues, the organized and unstructured information is joined in social insurance field for successfully anticipating the sicknesses [2].

2.1. Existing System

Prediction using traditional disease risk models usually involves a machine learning algorithm and especially a supervised learning algorithm. Here we divide the data as training and test data. In the test set, patients can be classified into groups of either high-risk or low-risk. These models are valuable in clinical situations.

With an incredible development in therapeutic information, gathering electronic wellbeing records (EHR) is progressively advantageous. In gathering information right off the bat introduced a bio motivated superior heterogeneous vehicular telematics worldview, to such an extent that the accumulation of versatile clients' wellbeing related information can be accomplished with the arrangement of cutting edge heterogeneous vehicular systems.

Chen et.al proposed a human services framework utilizing savvy attire for maintainable wellbeing checking.

Qiu et al had altogether considered the heterogeneous frameworks and accomplished the best outcomes for cost minimization on tree and straightforward way cases for heterogeneous frameworks. Patients' measurable data, test results and sickness history are recorded in the EHR, empowering us to recognize potential information driven answers for lessen the expenses of restorative contextual analyses.

Qiu et al. proposed a productive stream evaluating calculation for the tele wellbeing cloud framework and composed an information rationality convention for the PHR (Personal Health Record)- based dispersed framework.

Bates et al. proposed six utilizations of huge information in the field of social insurance.

Qiu et al. proposed an ideal enormous information sharing calculation to deal with the entangle informational collection in telehealth with cloud systems. One of the applications is to recognize high-hazard patients which can be used to decrease medicinal cost since high-chance patients frequently require costly human services [3].

Disadvantages of Existing System:

- In the existing system the data set is typically small, for patients and diseases with specific conditions.
- The pre-selected characteristics maybe not satisfy the changes in the disease and its influencing factors
- It provides low accuracy and it is more time consuming.
- It includes health risk assessment.

2.2 Proposed System

In the proposed system we combine the structured and unstructured data in healthcare field to assess the risk of disease. First, we used latent factor model to reconstruct the missing data from the medical records collected from a hospital. Second, by using statistical knowledge, we determine the major chronic diseases in the particular region. Third, to use structured data, we consult with hospital experts to extract useful features. For unstructured text data, we select the features automatically using CNN algorithm. Finally, we propose a novel CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm for disease prediction using both structured and unstructured data.

Advantages of Proposed System:

It provides higher accuracy.

- We leverage not only the structured data but also the text data of patients based on the proposed CNN-MDPR algorithm.

- We find that by combining these two data, the accuracy rate can reach high when we combine these two data.
- To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics.
- It also reduce the costs of medical case studies.

3. MODEL DESCRIPTION

The hospital dataset used in this paper contains real-life hospital data and the data are stored in the data center. A security access mechanism is created to protect the patient's privacy. The inpatient department data contains structured and unstructured text data. The structured data includes laboratory data and the patient's basic information such as the patient's age, gender and life habits etc. While the unstructured text data includes the patient's narration of his/her illness, the doctor's interrogation records and diagnosis etc. The goal of this study is to predict whether a patient has the chances of having any type of diseases from his medical history. In this we divide the data into training data and test data. For S-data, we use three conventional machine learning algorithms, i.e., Naive Bayesian (NB), K-nearest Neighbor (KNN), and Decision Tree (DT) algorithm to predict the risk of disease. For T-data, CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm is proposed to predict the risk of disease. For S&T data, we predict the risk of disease by the use of CNN-MDRP.

4. METHODS DESCRIPTION

In this paper we use three algorithms for predicting the diseases. One is KNN, second is Naïve Bayesian and third is Decision tree.

Data Imputation.

For patient's examination data, there will be a large number of missing data due to human error. Thus, we need to fill the structured data. Before data imputation, we first identify incomplete medical data and then modify or delete them to improve the data quality. Then, we use data integration for data pre-processing. We can integrate the medical data to guarantee data atomicity. For data imputation, we use the latent factor model which is presented to explain the observable variables in terms of the latent variables.

KNN Algorithm.

KNN is a sort of occurrence based learning in which the capacity is just privately approximated and all figuring's are put off until order and investigate the best information of the illness procedure. The KNN calculation is one of the least complex of all machine learning calculations. For both classification and relapse, a helpful technique might be to weight the neighbors' help so that the closer neighbors contribute more to the normal than the more far off ones. For instance, a typical weighting plan is to offer weight to each neighbor, which is the separation to the neighbor. The neighbors are involved from an arrangement of things for which the class or the protest property estimation (for the KNN relapse) is known. This can be considered as the preparation set for the calculation, albeit no unequivocal preparing step is required. An exceptional element of the k-NN calculation is that it is delicate to the nearby structure of the information. The calculation ought not to be mistaken for k-implies, another prevalent machine learning strategy.

Naive Bayesian.

The finding of a medicinal condition depends on the side effects, physical examination and restorative history of a patient. There have been numerous situations where treatment for an infection has not been done precisely because of absence of legitimate investigation and obliviousness of various imperative components. To make the fundamentals of good wellbeing hones open to everybody, a Naïve Bayes approach for Disease Diagnosis is proposed here. This administration plans to help patients and specialists by directing them to anticipate conceivable illnesses utilizing the side effects gave by the client. Guileless Bayes classifier is utilized for characterization of information. The framework contains a manifestation informational collection which is additionally sorted as Name, Attributes, and Record information. In light of the positioning of every side effect gave, the analyzed outcome recommends the likelihood weight of every illness. It additionally gives the detail of recognized sicknesses including the cure, counteractive action and the conceivable treatment.

Decision Tree.

In medicinal choice like arrangement, diagnosing there are numerous circumstances where choice must be made successfully and dependably. Reasonable straightforward basic leadership models with the likelihood of programmed learning are the most fitting for performing such undertakings. Choice trees are a solid and successful basic leadership procedure that furnish high grouping exactness with a straightforward portrayal of accumulated learning and they have been utilized as a part of various zones of restorative basic leadership.

CNN based Multimodal Disease Risk Prediction (CNN-MDRP) Algorithm.

CNN-UDRP only uses the text data to predict whether the patient has disease or not. As for both structured and unstructured text data, we design a CNN-MDRP algorithm based on CNN-UDRP. The processing of text data is similar with CNN-UDRP and computation methods are also similar with CNN-UDRP algorithm.

5. PROPOSED SYSTEM ARCHITECTURE

A new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm is proposed using structured and unstructured data from hospital. None of the existing work focused on both data types in the area of medical. Various researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification, rather than the previously selected characteristics.

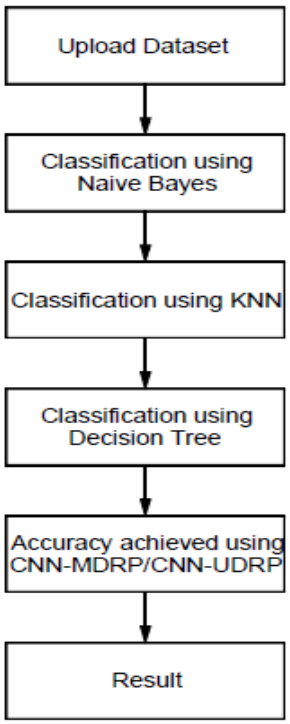

Fig : Proposed System Architecture

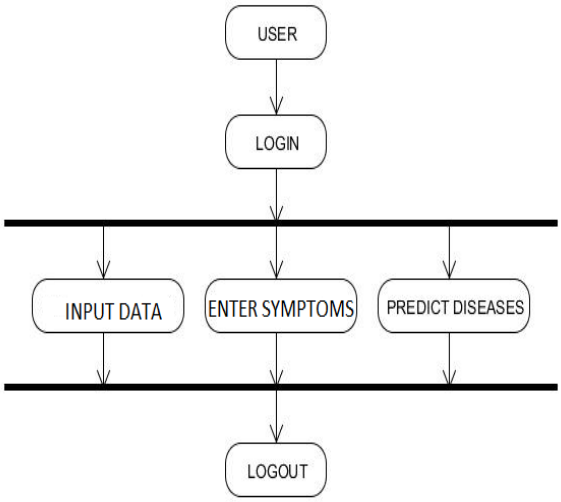

Fig: System Design of User

6. CONCLUSION

In this paper, we propose a machine learning and new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital for effective prediction of diseases. Existing work is not focused on both data types in the area of healthcare. Compared to several typical prediction algorithms, the proposed algorithm accuracy prediction reaches 94.8% than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm.

REFERENCES

[1] Anand Borad, Healthcare and Machine Learning: The Future with Possibilities Jan. 18.

[2] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, "Wearable 2.0: Enable human-cloud integration in next generation healthcare system, "IEEE Commun., vol. 55, no. 1, pp. 54_61, Jan. 2017.

[3] P. Groves, B. Kayyali, D. Knott, and S. van Kuiken, The Big Data Revolution in Healthcare: Accelerating Value and Innovation. USA: Center for US Health System Reform Business Technology Office, 2016.

[4] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction, "Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070_1093, 2015.

[5] S. Bandyopadhyayet al., "Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data", Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1033_1069, 2015.

BIOGRAPHIES:

Harini D K, pursuing M.Tech in computer science & Engineering, Vidyavardhaka college of Engineering, Mysuru, Karnataka.

Natesh M, Associate Professor in Department of computer science & Engineering, Vidyavardhaka college of Engineering, Mysuru, Karnataka.