ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

Predictive Analysis of Diseases Using Machine Learning and Big Data

Sireesha Kilaru[1], D. Anupama[2]

M.Tech Scholar, Dept. of CSE, Sai Tirumala NVR Engineering Collage, Jonnalagadda, Andhrapradesh, India[1]

Professor, Dept. of CSE, Tirumala NVR Engineering Collage, Jonnalagadda, Andhrapradesh, India [2]

ABSTRACT: Today we must accept the truth that machine learning and big data are boon to healthcare industry, and many research works are already started to reduce the complications in biomedical and healthcare fields. In order to get the maximum benefit from these two sectors we need quality data as input. This can be obtained by the healthcare historical databases. Incomplete data may not predict the result accurately; to avoid this problem our proposed system will use machine learning algorithms to predict the accurate disease in the onset of chronic disease outbreaks in disease-frequent communities. For this we have collected a regional hospital data. We can use latent factor model to achieve incomplete data. The proposed system will use neuronal network based multimodal disease risk prediction (CNN-MDRP) algorithm that uses structured and unstructured data from the hospital. To the best of my knowledge, none of the existing work focused on the both types of data in the area of healthcare big data analysis .Our proposed algorithm will have a prediction accuracy of 94.8% with a convergence rate faster than CNN-based unimodal disease risk prediction algorithm (CNN-UDRP).

KEYWORDS: Machine Learning Algorithms, Healthcare Industry, Historical Databases, Disease-frequent Communities, CNN-MDRP, CNN-UDRP, and Prediction Accuracy.

I. INTRODUCTION

Early detection is the key of success in biomedical and healthcare communities. The accurate analysis of medical data will predict accurate results. In addition different regions have unique characteristics of certain regional diseases, which can weaken the prediction of disease outbreaks. This paper will rationalize the machine learning algorithms to effectively predict the onset of chronic disease outbreaks in disease-frequent communities. Big data is an emerging field and looking forward to resolve all kind of analytical problems. Normally data will have three properties velocity, volume and variety, big data is the collection of data which will preserve the data properties to predict the disease. Most of the existing work will process the structured data and our proposed system will cover the unstructured data as well by using convolution neural networks (CNN). CNN is made up of neurons each neuron will receive an input and provide different outputs.

The existing work covered structured data but unstructured data is not taken which lead to inaccurate results we have to consider the following scenarios: How to collect the missing data, How to determine the regional disease characteristics, How to maintain the region climatic conditions, How to preserve the patient data, to get the accurate prediction. For this we are using CNN algorithm. Using CNN-MDRP we can process both structured and unstructured data types. This improved the prediction accuracy.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

II. OBJECTIVES

1. Predictive analysis of diseases using structured data can be done by the traditional machine learning algorithms like Naive Bayesian and SVM.
2. Predictive analysis of diseases using unstructured data can be done by using CNN-MDRP algorithm. This also improved the accuracy.

III. SYSTEM ARCHITECTURE

The proposed system architecture can be shown as below:

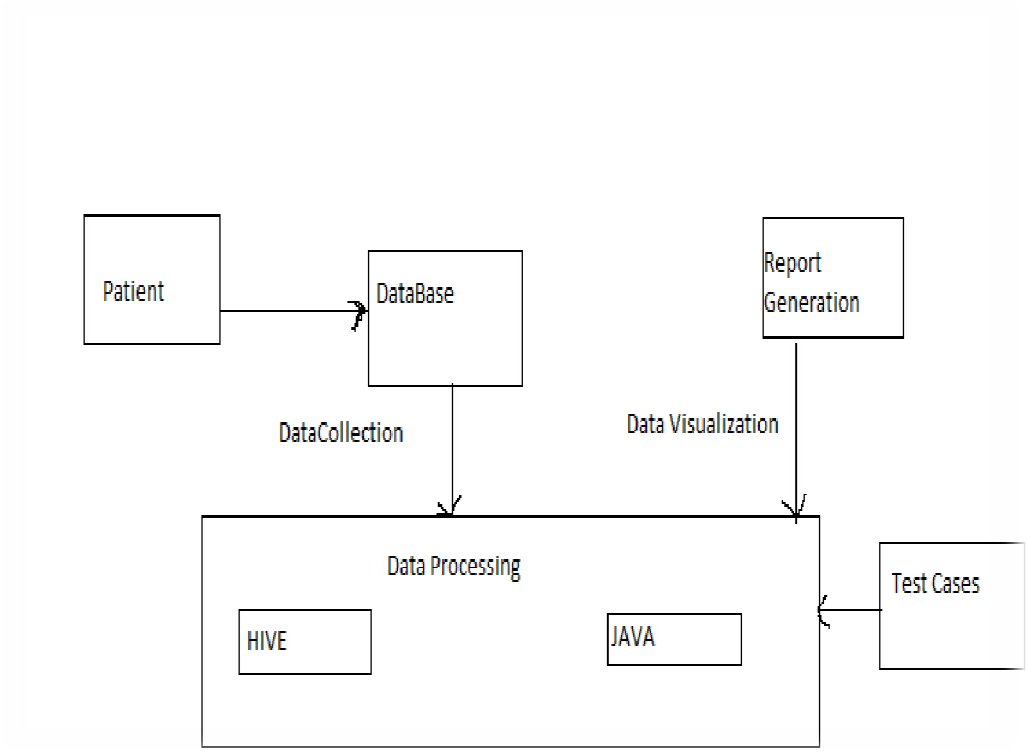


Fig 1: Architecture of the proposed system

In the proposed system we mainly use machine learning algorithm for prediction and hadoop HIVE and JAVA for processing the job. We can explain those two in shot as below:

MACHINE LEARNING: We can explicitly tell that machine learning is an application of artificial intelligence. Machine Learning advantage is it will learn from the previous experience and can be used in future without programming. This focuses on the development of computer programs that can access data and use it learn for themselves. Machine Learning delivers several methods and approaches that can help resolving analytic and predictive problems in medicinal areas Machine Learning is also being used for statistics examination, such as discovery of proportions in the data by rightly commerce with flawed data, clarification of incessant data used in the Strenuous Care Unit, and brainy troubling subsequent in real and ordered nursing. In our proposed system we used machine learning

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

algorithms to predict the patient's disease with the help of input given by the patient and by analysing the previous data. ML learns the data and provides the result.

MAP REDUCE: Map Reduce is a tool in big data which have the processing technique and a program model for distributed computing based on java. The Map Reduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. Hadoop Map Reduce is used in our project in order to identify the symptoms of each disease by taking from an input file and creating the key value pairs whereas the key is disease and values are symptoms.

REPORT: The report generated by the proposed system will take patient symptoms as input and this input can be provided as a csv file. This input is compared with the existing records of patients and database of disease symptoms and specific disease is identified and this can be confirmed by the test which is specified in the database, and as final step the patient is provided with a suggestion of specialist for that disease, which is present in the database.

EXISTING SYSTEM: The existing system is designed in such a way that only structured data from the hospitals and other medical lab data sets are taken as input and can be processed using CNN-UDRP algorithm to predict the disease and this produced the results with very less accuracy. And the drawback in this existing system is it is not considering the external factors at the time of input which is mentioned as unstructured data. So, the prediction possibility is only 60%. And there is no proper doctor profiles database to suggest the patient for later treatment. To cover all these problems we have created a proposed system.

PROPOSED SYSTEM : The proposed system will collect all structured and unstructured data from the regional healthcare centres and laboratories to analyse the diseases which are frequently occurring in certain regions and it will also maintain the specialists profiles for those diseases and when there is a scenario where the patient will provide with symptoms as input and our proposed system will use CNN-MDRP algorithm and big data HIVE to compare and extract the results and the maximum prediction accuracy have improved to 94% in the proposed system. This also reduced the cost when compared to existing system. The comparison and extraction speed is also increased because of using big data.

FLOW CHART: This flow chart represents the overall functioning of our proposed system:

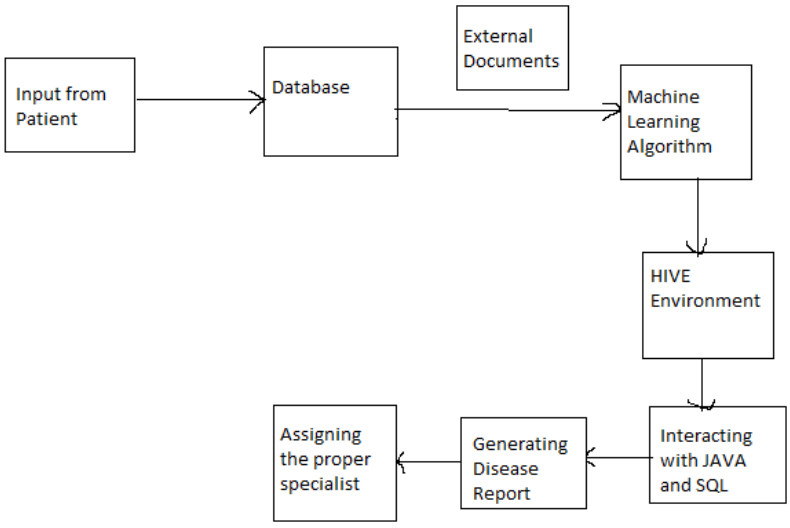


Fig 2: flow chart of the proposed system

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

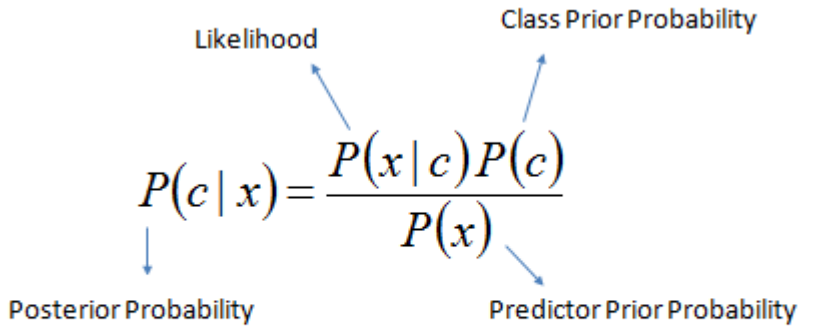
(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

The above flow chart will represent the way medical practitioner is going to predict the disease by providing the patient symptoms and get the related doctor profile for treatment. Prediction can be done by Bayesian algorithm and disease classification can be done using CNN-MDRP algorithm.

SYSTEM ANALYSIS: The system analysis shows how the prediction of the disease can be done.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig 3: Naives Bayesian algorithm for proposed system

The above diagram shows the inputs to Naive Bayesian algorithm.

- $P(c|x)$ 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*.

Working of the system: The overall working of the system is shown in the below diagram.
The below diagram contains the roles of all persons involved in the system and their tasks. The persons involved are:
Administrator: Admin is responsible for the interaction with the system by getting patient disease symptoms and provide it to system to apply on the algorithms.
Medical Practitioner: Medical practitioner is the next person to doctor who is responsible for generating graph and selecting the respective specialist for the patient.
Patient: The patient is the use who is using the system to get benefit. He will pass the symptoms he is facing to the system and obtain proper graph of disease and to which specialist to meet.
Proposed System: The proposed system is machine learning algorithm and Big Data code to process the patient data. The machine learning algorithm is CNN-MDRP and HIVE of hadoop. For data processing in HIVE synchronization is done through JAVA.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

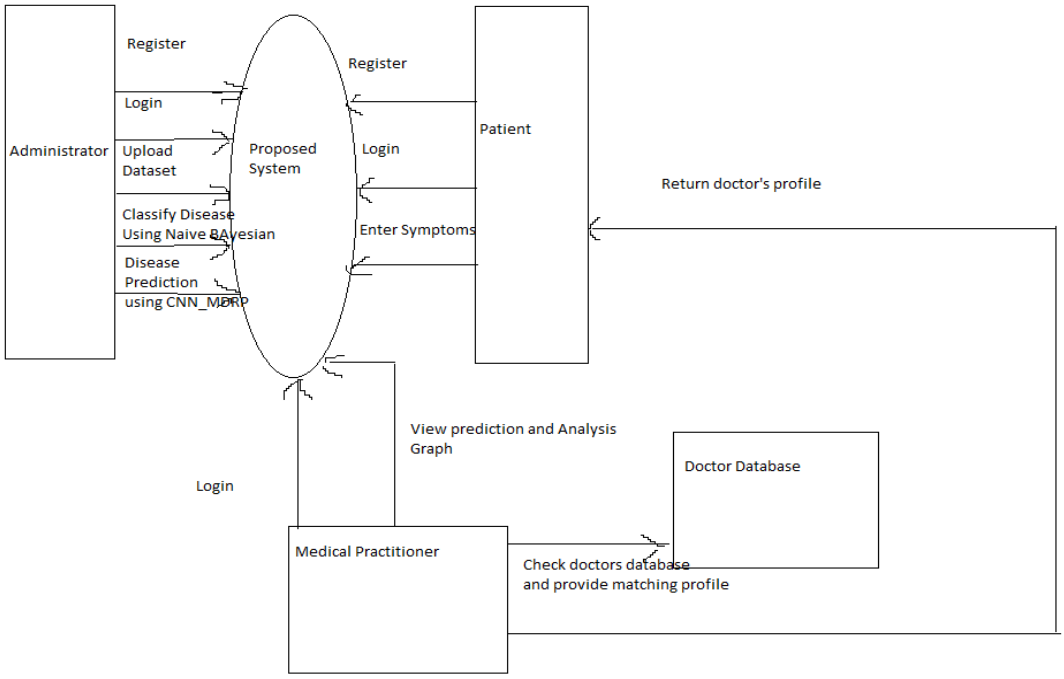


Fig 4: Data flow of the proposed system

IV. EXPERIMENTAL RESULTS

The experimental results for the given input can be shown as follows:

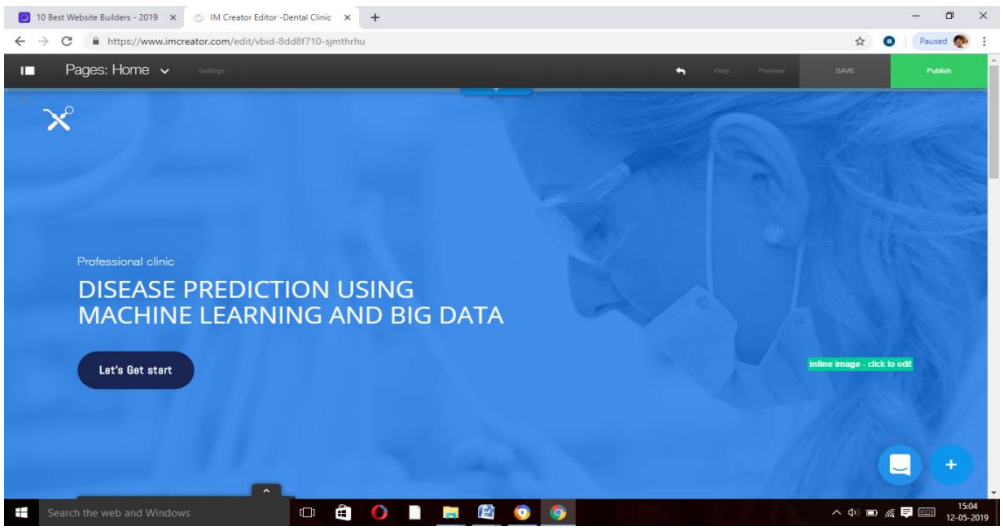


Table 1: input values for Disease Prediction System

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

Patient_ID	Input
P_101	Nausea, Vomiting, Blood Vomit.
P_102	Skin irritation, rashes, allergy.
P_103	Pain in the lower abdomen and bladder .
P_104	Knee and joint pains.

The above mentioned inputs are given to system by patient and then the whole system process is done and the expected result can be as follows.

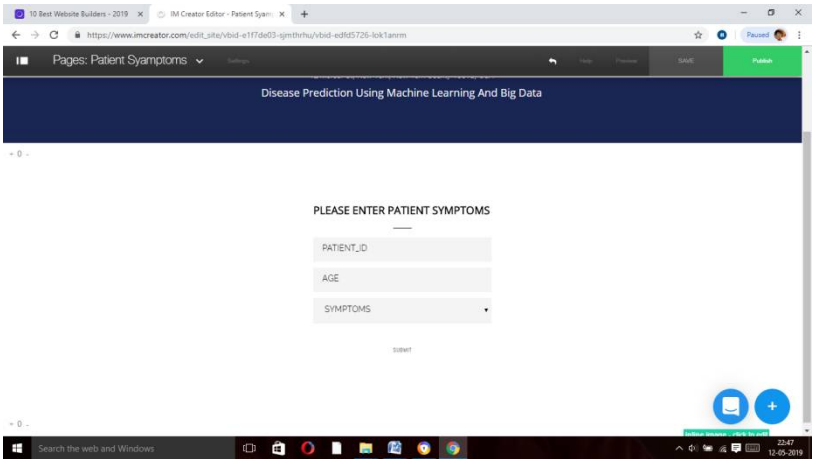


Fig 5: Input through GUI

The above fig shows input specification to the system. All patient symptoms are collected to Hadoop database and are processed further.

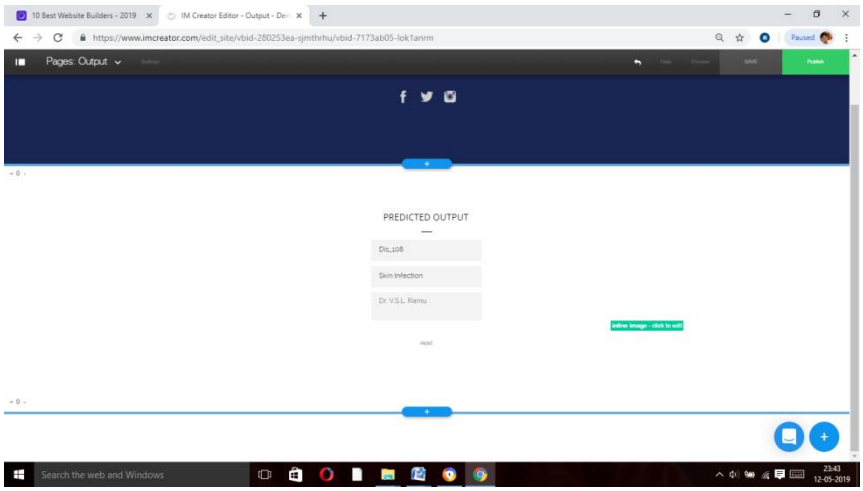


Fig 6: predicted output through GUI.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

The above figure will illustrate the fetched results from hadooop file system. We can obtain by running hadoop HIVE job and saving the obtained result to dataset and based on disease_id the result set is obtained.

Result of the proposed system: The proposed system will also prove that CNN-MDRP is better in performance when compared to CNN-UDRP.

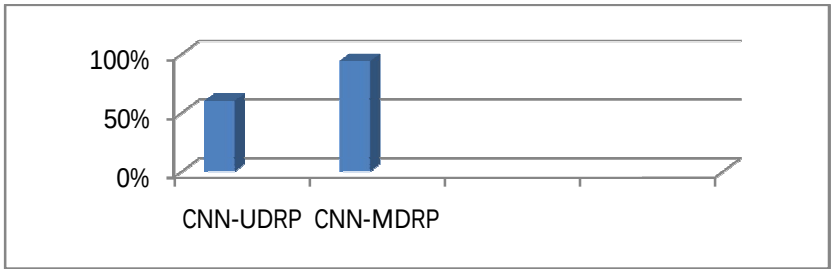


Fig 7: Comparison of CNN-URDP and CNN-MDRP algorithms.

System Configuration: We need the below configurations to run our project.

1. Python
2. Java
3. Hadoop environment
4. HIVE environment
5. Eclipse software

To implement the above job we need a Pentium i4 or above processor with 4GB or high RAM and 200GB of disk space.

V. CONCLUSION

In this paper we are going to use CNN-MDRP machine learning algorithm to process unstructured data which is not possible for the existing system using CNN-URDP algorithm. One more enhancement for the existing system is it is shown in the perspective of the doctor in the existing system but in our proposed system we are simplifying in such a way that all prediction and testing can be done in medical representative level and for last clearance we are opting for concerned specialist which is specified in the database.

REFERENCES

[1]. M. Thakur and B. Singh, “Simulink Modal of Triple-Junction Solar Cell and MPPT Based on Incremental Conductance Algorithm for PV System,” vol. 5, no. 9, pp. 92–95, 2015.
[2]. V. K. P. K, C. A. Asha, and M. K. Sreenivasan, “design , simulation and hardware implementation of efficient solar power converter with high mpp tracking accuracy for dc microgrid applications,” pp. 380–387, 2014.
[3]. K. Supreeth, K. S. Sundar, and D. Balamurugan, “Performance Evaluation & Simulation of Solar Power System,” vol. 2, no. 7, 2014.
[4]. J. Park, H. Kim, Y. Cho, and C. Shin, “Simple Modeling and Simulation of Photovoltaic Panels Using Matlab / Simulink Modeling of Photovoltaic Module,” vol. 73, no. Fgcn, pp. 147–155, 2014.
[5]. A. Shukla, M. Khare, and K. N. Shukla, “Modeling and Simulation of Solar PV Module on MATLAB / Simulink,” pp. 18516–18527, 2015.
[6]. E. G. S, P. S. Dhivya, and T. Sivaprakasam, “Solar Powered High Efficient Dual Buck Converter for Battery Charging,” pp. 448–452,2013.
[7]. J. Patel and G. Sharma, “modeling and simulation of solar photovoltaic module using matlab / simulink,” pp. 225–228, 2013.
[8]. S. S. Mohammed-, “Modeling and Simulation of Photovoltaic module using MATLAB / Simulink,” vol. 2, no. 5, 2011.