Designing Disease Prediction Model Using Machine Learning Approach

[1]Dhiraj Dahiwade,[2]Prof. Gajanan Patle, [2]Prof. Ektaa Meshram

[1]PG Scholar, Department of Computer Science Engineering, Abha-Gaikwad Patil College of Engineering, Nagpur, Maharashtra, India.

[2]Assistant Professor, Department of Computer Science Engineering, Abha-Gaikwad Patil College of Engineering, Nagpur, Maharashtra, India.

Abstract— **Now-a-days, people face various diseases due to the environmental condition and their living habits. So the prediction of disease at earlier stage becomes important task. But the accurate prediction on the basis of symptoms becomes too difficult for doctor. The correct prediction of disease is the most challenging task. To overcome this problem data mining plays an important role to predict the disease. Medical science has large amount of data growth per year. Due to increase amount of data growth in medical and healthcare field the accurate analysis on medical data which has been benefits from early patient care. With the help of disease data, data mining finds hidden pattern information in the huge amount of medical data. We proposed general disease prediction based on symptoms of the patient. For the disease prediction, we use K-Nearest Neighbor (KNN) and Convolutional neural network (CNN) machine learning algorithm for accurate prediction of disease. For disease prediction required disease symptoms dataset. In this general disease prediction the living habits of person and checkup information consider for the accurate prediction. The accuracy of general disease prediction by using CNN is 84.5% which is more than KNN algorithm. And the time and the memory requirement is also more in KNN than CNN. After general disease prediction, this system able to gives the risk associated with general disease which is lower risk of general disease or higher.**

Index Terms: CNN, KNN, Machine learning, Disease Prediction.

I. INTRODUCTION

Artificial Intelligence made computer more intelligent and can enable the computer to think. AI study consider machine learning as subfield in numerous research work. Different analysts feel that without learning, insight can't be created. There are numerous kinds of Machine Learning Techniques like Unsupervised, Semi Supervised, Supervised, Reinforcement, Evolutionary Learning and Deep Learning. These learnings are used to classify huge data very fastly. So we use K-Nearest Neighbor (KNN) and Convolutional neural network (CNN) machine learning algorithm for fast classification of big data and accurate prediction of disease. Because medical data is increasing day by day so usage of that for predicting correct disease is crucial task but processing big data is very crucial in general so data mining plays very important role and classification of large dataset using machine learning becomes so easy.

It is critical to comprehend the accurate diagnosis of patients by clinical examination and evaluation. For compelling determination decision support systems that depend on computer may assume an indispensable job. Health care field creates enormous information about clinical evaluation, report in regards to patient, cure, subsequent meet-ups, medicine and so forth. It is intricate to orchestrate appropriately. Quality of the data association has been influenced due to improper management of the information. . Upgrade in the measure of data needs some legitimate way to concentrate and process information viably and efficiently. One of the many machine-learning applications is utilized to construct such classifier that can separate the data based on their characteristics. Data set is partitioned into two or more than two classes. Such classifiers are utilized for medical data investigation and disease prediction.

Today machine learning is present everywhere so that without knowing it, one can possibly use it many times a day. CNN uses both the structured and unstructured data of a hospital to do classification. While other machine learning algorithms only work on structured data and time required for computation is high also they are lazy because they store entire data as a training dataset and uses complex method for calculation.

The section I explains the Introduction of general disease prediction using classification method such as KNN and CNN. Section II presents the literature review of existing systems and Section III present proposed system implementation details Section IV presents experimental analysis, results and discussion of proposed system. Section V concludes our proposed system. While at the end list of references paper are presented.

II. LITERATURE REVIEW

M. Chen proposed [1] a new CNN based multimodal disease risk prediction algorithm by using structured and unstructured data of hospital. M. Chen ,Y. Hao, K. Hwang, L. Wang, and L. Wang invented disease prediction system for the numerous regions. They performed disease prediction on three diseases like diabetics, cerebral infraction and heart disease. The disease prediction is carried out on structured data. Prediction of heart disease, diabetes and cerebral infraction is carried out by using different machine learning algorithm like naïve bayes, Decision tree and KNN algorithm. The result of Decision tree algorithm is better than Naïve bayes and KNN algorithm. Also, they predict that whether a patient experiences from the high risk of cerebral infarction or low risk of cerebral infarction. For the risk prediction of cerebral infraction, they utilized CNN based multimodel disease risk prediction on text data. The accuracy comparison takes place between CNN based unimodel disease risk predictions against CNN based multimodel disease risk prediction algorithm. The accuracy of disease prediction reaches up to the 94.8% with faster speed than CNN based unimodal disease risk prediction algorithm. The CNN based multimodel disease risk prediction algorithm steps is similar as that of the CNN-UDRP algorithm only the testing steps consist of two additional steps. This paper work on both the type of dataset like structured and unstructured data. Author worked on unstructured data. While previous work only based on structured data, none of the author worked on unstructured and semi- structured data. But this paper depends on structured as well as unstructured data.

B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang [2] designed the Alzheimer disease risk prediction system with the help of EHR data of the patient. Here they utilized active learning context to solve a real problem suffered by the patient. In this the active patient risk model was build. For that active risk prediction algorithm is utilized the risk of Alzheimer disease.

IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn [3] proposed wearable 2.0 system in which design smart washable clothing that improves the QoE and QoS of the next-generation healthcare system. Chen designed new IoT based data collection system. In that new sensor based smart washable cloth invented. By the used of this cloth, doctor captured the patient physiological condition. And with the help of the physiological data further analysis happen. In this inversion of washable smart cloth mainly consists of multiple sensors, wires and electrode with the help of this component user can able to collect the physiological condition of patient as well as emotional health status information by the used of cloud based system. With the help of this cloth, it captured the physiological condition of the patient. And for the analysis purpose, this data is used. Discussed the issues which are facing while designed the wearable 2.0 architecture. The issues in existing system consist of physiological data collection, negative psychological effects, anti-wireless for body area networking and Sustainable big physiological data collection etc. The multiple operations performed on files like analysis on data, monitoring and prediction. Again author classify the functional components of the smart clothing representing Wearable 2.0 into the following categories like sensors Integration, electrical-cable-based networking, digital modules. In this, there are many applications discussed like chronic disease monitoring, elderly people care, emotion care etc.

Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri [4] designed cloud-based health –Cps system in which manages the huge amount of biomedical data. Y. Zhang discussed large amount of data growth in the medical field. The data is created within the less amount of time and the characteristic of data is stored in different format so this is what the problem related to the big data. In this designed the health-Cps system in that two technologies prefer one is cloud and second one is big data technology. This system performed numerous operations on cloud-like data analysis, monitoring and prediction of data. With the help of this system, a person gets more information about how to handle and manage the huge amount of biomedical data in the cloud. The three layers consider in the system data collection layer, data management layer and data- oriented layer. The data collection layer stored data in the particular standard format. The data management layer used for distributed storage and parallel computing. By this system multiple operations are performed with the help of Health-cps system. Also, the many services related to healthcare know by this system.

L. Qiu, K. Gai, and M. Qiu in [5] proposed telehealth system and discussed how to handle a large amount of hospital data in the cloud. This paper author proposed advance in the telehealth system, which is mainly based on the sharing data among all the telehealth services over the cloud. But the data sharing on the cloud facing lots of issues like network capacity and virtual machine switches. In this proposed the data sharing on cloud approach for the better sharing of data through the data sharing concepts. Here designed the optimal method of telehealth sharing model. By this model, author focus on transmission probability, network capabilities and timing constraints. For this author invented new optimal big data sharing algorithm. By this algorithm, users get the optimal solution of handling biomedical data.

Ajinkya Kunjir, Harshal Sawant, Nuzhat F.Shaikh [6] proposed a best clinical decision-making system which predicts the disease on the basis of historical data of patients. In this predicted multiple diseases and unseen pattern of patient condition. Designed a best clinical decision- making system used for the accurate disease prediction on the historical data. In that also determined multiple diseases concept and unseen pattern. For the visualization purpose in this used 2D/3D graph and pie Charts.And 2D/3D graph and pie charts designed for visualization purpose.

S.Leoni Sharmila, C.Dharuman and P.Venkatesan [13] gives a comparative study of different machine learning technique such Fuzzy logic, Fuzzy Neural Network and decision tree. They consider liver data set to classify and do study comparatively. According to study Fuzzy Neutral Network gives 91% accuracy for classification in liver disease dataset than other machine learning algorithm. Author used Simplified Fuzzy ARTMAP in varied nature of application domains and is competent to perform classification very efficiently and giving very high performances.

Author have concluded that machine learning algorithms such as Naive Bayes and Apriori [14] are highly useful for disease diagnosis on the given data set. Here small volume data used for prediction like symptoms or previous knowledge obtained from the physical diagnosis. Limitation of this paper they could not consider large dataset, now a days medical data is growing so needs to classify that and classification of that data is challenging.

Shraddha Subhash Shirsath [15] proposed 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. Here they consider big data.

III. SYSTEM ARCHITECTURE

A. Architecture Overview

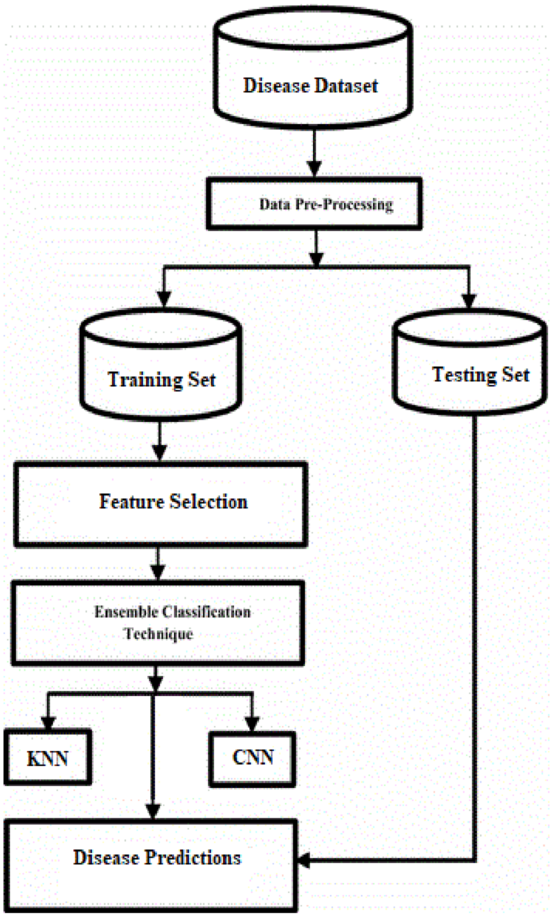

Fig 1. System Architecture

Initially we take disease dataset from UCI machine learning website and that is in the form of disease list with its symptoms. After that preprocessing is performed on that dataset for cleaning that is removing comma, punctuations and white places. And that is used as training dataset. After that feature extracted and selected. Then we classify that data using classification techniques such as KNN and CNN. Based on machine learning we can predict accurate disease.

B. Algorithms Used

1) K-Nearest neighbor (KNN)

Common Distance Metrics:

Euclidean distance (continuous distribution):

$d(a,b) = \sqrt{}\sum(a_i - b_i)^2$

Hamming distance (overlap metric):
bat (distance = 1)
toned (distance = 3)

Discrete Metric (boolean metric):
if x = y then d(x,y) = 0. Otherwise, d(x,y) = 1

Determine the class from *k* nearest neighbor list
Take the majority vote of class labels among the k-nearest neighbors

Weighted factor:
w =1/d (generalized linear interpolation) or $1/d^2$

2) *Convolutional neural network (CNN)*

Step 1: The dataset is converted into the vector form.
Step 2: Then word embedding carried out which adopt zero values to fill the data. The output of word embedding is convolutional layer.
Step 3: This Convolutional layer taken as input to pooling layer and we perform max pooling operation on convolutional layer.
Step 4: In Max pooling the dataset convert into fixed length vector form. Pooling layer is connected with the full connected neural network.
Step 5: The full connection layer connected to the classifier that is softmax classifier.

IV. RESULT AND DISCUSSIONS

A. Experimental Setup

1) All the experimental cases are implemented in Java in congestion with Netbeans tools and MySql as backend, algorithms and strategies, and the competing classification approach along with various feature extraction technique, and run in environment with System having configuration of Intel Core i5-6200U, 2.30 GHz Windows 10 (64 bit) machine with 8GB of RAM

2) Dataset

Patient disease dataset downloaded from UCI machine learning website.

B. Comparison Results

This section presents the performance of the KNN and CNN classification algorithms in terms of time required and accuracy. Fig 2 Shows accuracy Comparison of KNN and CNN algorithms for various Threshold. X-axis shows Algorithm & Y-axis shows accuracy in %. CNN gives more accurate disease prediction than KNN.

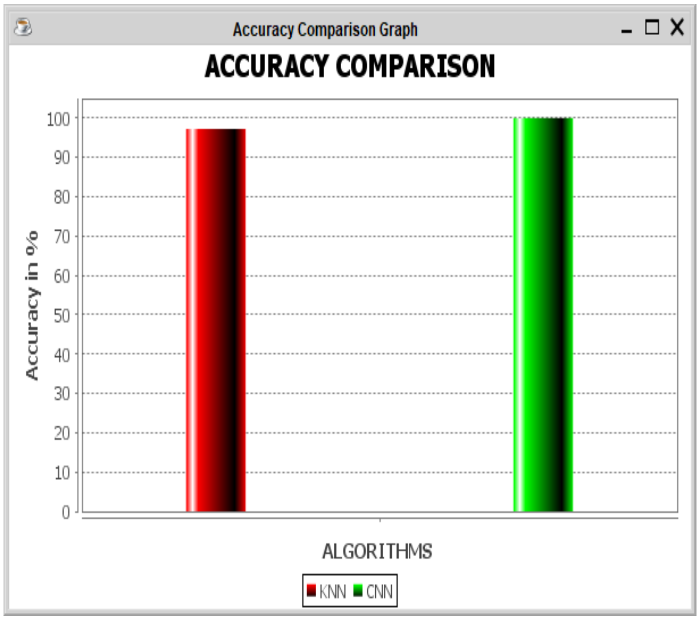

Fig. 2: Accuracy Comparison Graph

Fig 3 shows the Time comparison of KNN and CNN algorithms for various size. The X-axis shows algorithms and Y- axis shows Time in ms. The CNN takes less time than KNN for classifying large dataset.

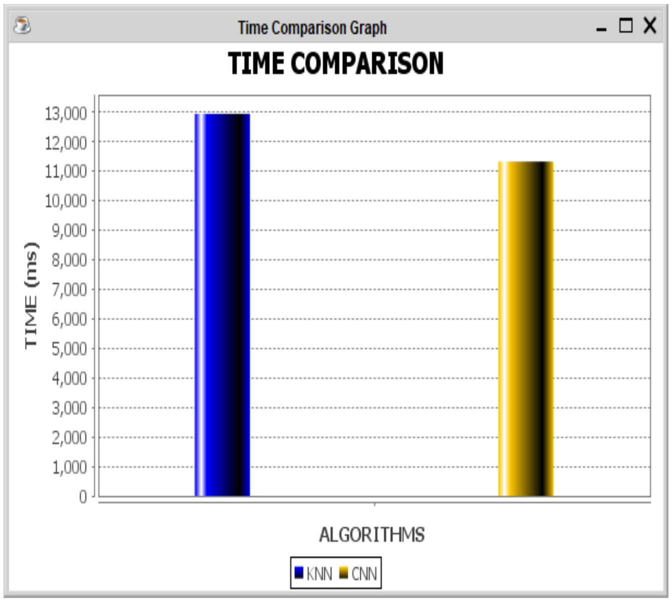

Fig. 3: Time Comparison Graph

V. CONCLUSION

We proposed general disease prediction system based on machine learning algorithm. We utilized KNN and CNN algorithms to classify patient data because today medical data growing very vastly and that needs to process existed data for predicting exact disease based on symptoms. We got accurate general disease risk prediction as output, by giving the input as patients record which help us to understand the level of disease risk prediction. Because of this system may leads in low time consumption and minimal cost possible for disease prediction and risk prediction. We compare the results between KNN and CNN algorithm in terms of accuracy and time and the accuracy of CNN algorithm which is more than KNN algorithm and time required for classification for CNN is less than KNN. So we can say CNN is better than KNN in terms of accuracy and time.

REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang,"Disease prediction by machine learning over big data from healthcare communities", ," *IEEE Access,* vol. 5, no. 1, pp. 8869–8879, 2017.

[2] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," *Springer Data Mining Knowl. Discovery,* vol. 29, no. 4, pp. 1070–1093, 2015.

[3] IM. 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.

[4] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "HealthCPS: Healthcare cyberphysical system assisted by cloud and big data," *IEEE Syst.* J., vol. 11, no. 1, pp. 88–95, Mar. 2017.

[5] L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for telehealth in cloud computing," *in Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud),* Nov. 2016, pp. 184–189.

[6] Disease and symptoms Dataset *–www.github.com.*

[7] Heart disease Dataset-*WWW.UCI Repository. com*

[8] Ajinkya Kunjir, Harshal Sawant, Nuzhat F.Shaikh, "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare," *in IEEE big data analytics and computational intelligence*, Oct 2017 pp.2325.

[9] Shanthi Mendis, Pekka Puska, Bo Norrving, World Health Organization (2011), Global Atlas on Cardiovascular Disease Prevention and Control, PP. 3– 18. ISBN 978-92-4-156437-3.

[10] Amin, S.U.; Agarwal, K.; Beg, R., "Genetic neural network based data mining in prediction of heart disease using risk factors*", IEEE Conference on Information & Communication Technologies (ICT),* vol., no.,pp.1227-31,11-12 April 2013.

[11] Palaniappan S, Awang R, "Intelligent heart disease prediction System using data mining techniques," *IEEE/ACS International Conference on Computer Systems and Applications*, AICCSA 2008., vol., no., pp.108115, March 31 2008-April 4 2008.

[12] B. Nithya , Dr. V. Ilango Professor, "Predictive Analytics in Health Care Using Machine Learning Tools and Techniques," *International Conference on Intelligent Computing and Control Systems*,2017.

[13] S.Leoni Sharmila, C.Dharuman and P.Venkatesan "Disease Classification Using Machine Learning Algorithms - A Comparative Study", *International Journal of Pure and Applied Mathematics* Volume 114 No. 6 2017, 1-10

[14] Allen Daniel Sunny1, Sajal Kulshreshtha, Satyam Singh3, Srinabh, Mr. Mohan Ba, Dr. Sarojadevi H " Disease Diagnosis System By Exploring Machine Learning Algorithms", *International Journal of Innovations in Engineering and Technology (IJIET)* Volume 10 Issue 2 May 2018.

[15] Shraddha Subhash Shirsath "Disease Prediction Using Machine Learning Over Big Data" International Journal of Innovative Research in Science, Vol. 7, Issue 6, June 2018.