IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 27-32

Designing a Disease Prediction Model using Machine Learning

Ms.Jyoti Chandrashekhar Bambal[1], Prof. Roshani B. Talmale[2]

[1]*M-Tech student Department of Computer Science and Engineering,*
Tulsiramji Gaikwad-Patil College of Engineering & Technology, Nagpur, Maharashtra, India.
[2]*Assistant Professor Department of Computer Science and Technology*
Tulsiramji Gaikwad-Patil College of Engineering & Technology, Nagpur, Maharashtra, India.

Abstract: *Now a day, people face various diseases due to the environmental condition and living habits of them. So prediction of disease at earlier stage becomes important task. But the prediction on the basis of symptoms becomes too difficult for doctor. The correctly prediction of disease is most challenging task. To overcome this problem data mining plays an important and efficient way to predict the disease. Medical science has huge 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 large amount of medical data. We have designed the heart disease prediction system. We proposed multiple disease prediction based on symptoms of the patient. For the heart disease prediction, we used knn , naïve bayes machine learning algorithm for accurate prediction of disease. For disease prediction required disease symptoms dataset. Here we focused on heart disease prediction, because the heart disease is one of the leading causes of death among all other diseases. The heart disease prediction contains that whether the patient suffer from heart disease or not by using naïve bayes and KNN algorithm. In this heart disease prediction, the living habits of person and checkup information consider for the accurate prediction. The accuracy of heart disease prediction by using naïve bayes is 94.5% which is more than KNN algorithm. And the time and the memory requirement is also more in KNN than naïve bayes. After heart disease prediction, this system able to gives the risk associated with heart disease which is lower risk of heart disease or higher. For the risk prediction, we are using CNN algorithm.*

Index Terms: *Classification algorithm, machine learning, heart diseases prediction, data mining*

I. Introduction

Human face lots of problems related to the chronic disease. The main reason behind increase the chronic disease such as improper living habits, insufficient physical exercise, unhealthy diet, and irregular sleeping. 80% of people in the United States, spent more amount on the diagnosis of chronic disease. People give more aid for accurate prediction of disease.

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), Naïve Bayes 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 NB, 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 multimodal disorder risk prediction algorithm based on Convolutional Neural Network (CNN) by using organized and unorganized data of hospital. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang Discovered disease prediction system for various regions. They performed disease prediction on three different diseases such as diabetics, cerebral infraction and heart disease. The disease prediction is performed on organized data. Prediction of heart disease, diabetes and intellectual infraction is performed by using various machine learning algorithm like naïve bayes, Decision tree and KNN algorithm. The outcome of Decision tree algorithm performs better than KNN algorithm and Naïve bayes. Also, they predict that either a patient experience from the high risk of cerebral infarction or minimum risk of cerebral infarction. They used CNN based multimodal disease risk prediction on text data, for the risk prediction of cerebral infraction. The accuracy comparison takes place between CNN based unimodal disease risk predictions against CNN based multimodal disease risk prediction algorithm. The accuracy of disease prediction resulted up to the 94.8% with more fast speed than CNN based unimodal disease risk prediction algorithm. Step of similar as that of the CNN-UDRP algorithm the CNN based multimodal disease risk prediction algorithm step only the testing steps contains of two additional steps. Given paper work on both the type of dataset like organized and unorganized data. Author worked on unorganized data. While previous work only based on organized data, none of the author worked on unorganized and semi- organized data. But this system proposed work is depending on organized as well as unorganized data.

B. Qian, X. Wang, N. Cao, H. Li, and Y.- G. Jiang [2] planned the Alzheimer disease risk prediction system with the assistance of EHR information of the patient. Here they used active learning context to tackle a genuine issue endured by the patient. In this the risk model was construct. For that active risk prediction algorithm is used the risk of Alzheimer disease.

IM. Chen, Y. Mama, Y. Li, D. Wu, Y. Zhang, and C. Youn [3] proposed wearable 2.0 system in which configuration keen washable cloth that improves the QoE and QoS of the next-generation healthcare system. Chen structured new IoT based data collection system. In that new sensor based smart washable clothes created. By the utilized of this clothes, specialist caught the patient physiological condition. What's more, with the assistance of the physiological data analysis occur. In this reversal of washable smart cloth consisting of multiple sensor, wires and cathode with the assistance of this part component user can ready to gather the physiological state of patient as well as emotional health status information used of cloud based system. With the assistance of this material, it caught the physiological state of the patient. Also, for the examination reason, this information is utilized. Examined the issues which are confronting while designing wearable 2.0 architecture. The issues in existing system consist of physiological data gathering, negative mental impacts, anti-wireless for body zone networking and Sustainable big physiological data accumulation and so on. The numerous activities performed on records like examination on data, monitoring and prediction. Again author classify the functional components of the smart clothing representing Wearable 2.0 into sensors Integration, electrical-cable-based networking, digital modules. In this, there are numerous applications talked about 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 deals with the gigantic measure of biomedical data. Y. Zhang examined huge measure of information development in the medicinal field. The information is made inside the less measure of time and the normal for information is put away in various configuration so this is the thing that the issue identified with the big data. In this designed the Health-Cps system in that two advancements lean one is cloud and second one is big data technology. Cloud-like data analysis, monitoring and prediction of data. With the assistance of this system, an individual gets more data about how to deal with and deal with the enormous measure of biomedical information in the cloud. The three layers consider data collection layer, data management layer and data-oriented layer. The data gathering layer put away distributed storage and parallel computing. The data management layer used for distributed storage and parallel computing. By this framework various tasks are

performed with the assistance of Health-cps system, the many Health-cps systems. Related to healthcare know by this system.

L. Qiu, K. Gai, and M. Qiu in [5] proposed telehealth system and examined how to deal with a lot of hospital data in the cloud. This paper author proposed advance in the telehealth system, which is for the most part dependent on the sharing data among all the telehealth services over the cloud. Yet, the information sharing on the cloud confronting loads of issues like network capacity and virtual machine switches. In this proposed the data sharing on cloud approach for the better sharing of information through the data sharing ideas. Here planned the ideal strategy for telehealth sharing model. this model, author focus on transmission probability, network capabilities and timing constraints. For this creator concocted new big data sharing algorithm. By this calculation, clients get the ideal arrangement of handling biomedical data.

Ajinkya Kunjir, Harshal Sawant, Nuzhat F.Shaikh [6] proposed a best clinical decision-making system which predicts the disease based on historical data of patients. In this predicted various diseases and inconspicuous example of patient condition. Designed a best clinical decision- making system utilized for the exact disease prediction on the historical data. In that additionally decided various diseases concept and concealed example. For the perception reason in this used 2D/3D graph and pie Charts.And 2D/3D graph and pie charts representation reason.

S.Leoni Sharmila, C.Dharuman and P.Venkatesan [13] gives a similar investigation of various machine learning technique such Fuzzy logic, Fuzzy Neural Network and decision tree. They consider data set to classify and do study about nearly. As indicated by study Fuzzy Neutral Network gives 91% accuracy for classification in liver disease dataset than other machine learning algorithm. Author utilized Simplified Fuzzy ARTMAP in changed nature of application domains and is capable to perform classification all around productively and giving exceptionally superior performances.

Author have reasoned that machine learning algorithms for example, Naive Bayes and Apriori [14] are very valuable for disease diagnosis on the given data set. Here little volume data utilized for prediction like symptoms or past learning got from the physical diagnosis. Confinement of this paper they couldn't consider huge dataset, presently a day's medicinal data is developing so needs to classify that and classification of that information is challenging.

Shraddha Subhash Shirsath [15] proposed a CNN-MDRP algorithm for a disease prediction from an extensive volume of hospital's organized and unstructured information. Utilizing a machine learning algorithm (Neavi- Bayes) Existing algorithm CNN-UDRP just uses an organized information however in CNN-MDRP center around both organized and unstructured information the accuracy of disease prediction is more and quick when contrasted with the CNN-UDRP. Here they think about consider big data.

III. System Architecture

A. System Architecture

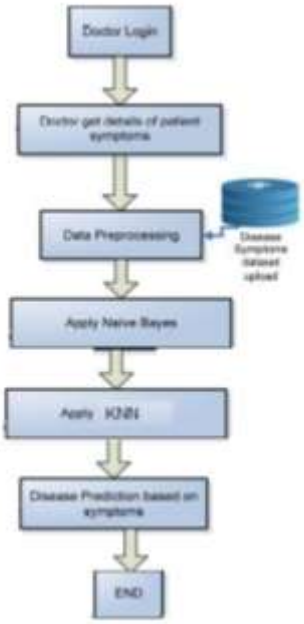

Fig 1. System Architecture

Fig1. Represent system architecture of proposed system. Initially we take disease dataset from UCI machine learning website and that is in the form of heart disease patient with different attributes and with its symptoms. After that preprocessing is performed on that dataset for cleaning that is removing unnecessary attributes and that is used as training dataset. After that feature extracted and selected. Then we classify that data using classification techniques such as NB, KNN and CNN. Based on machine learning we can predict accurate heart disease prediction.

B. Algorithms

1) Naïve Bayes Algorithm

Step 1: Consider the training set of tuples be the T. It contains n number of attributes they are X= (x1, x2, x3…… Xn) etc.

Step 2: Let K number of class for prediction (c1, c2, c3…. Cn). The classifier predicts that x will belong to that class who's having maximum posterior probability. If want to predict x values among two probability class that is p(ci/x) and p(cj/x). Then the values of p(ci/x) should be maximum, if we want that x values into the class.

Step 3: Create Likelihood table.

Step 4: Calculate posterior probability for each class by using naïve Bayesian equation.

Step 5: For prediction of X variable which is having high probability for the prediction outcomes.

Step 6: Stop

2) CNN Algorithm

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

C. Experimental Setup

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

D. Dataset Description

Patient disease dataset downloaded from UCI machine learning website.

E. Result

This section presents the performance of the KNN and NB classification algorithms in terms of time required and memory and other performance measures such as FP measure, precision, recall and accuracy.

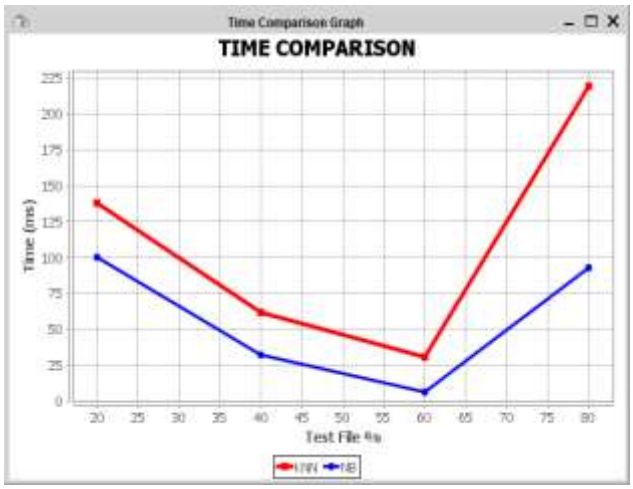

Fig 2. Time comparison line graph

Fig 2 Shows time comparison of KNN and NB algorithms for various Threshold. X-axis shows Algorithm & Y-axis shows time in ms. NBrequires less time than KNN.

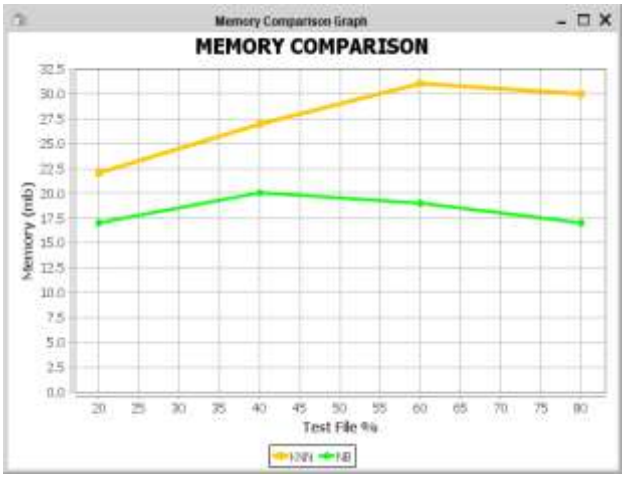

Fig 3. Memory comparison line graph

Fig 3 shows the Line graph of memory comparison of KNN and NB algorithms for various test file size. The X-axis shows test file size and Y- axis shows memory in bytes. The NB takes less memory than KNN for classifying large dataset.

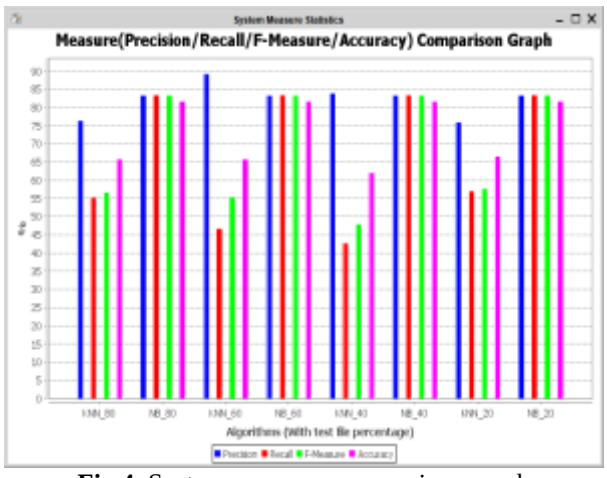

Fig 4. System measures comparison graph

Fig 4 shows the Bar graph of Algorithms measures (Precision/ recall/ f-measure/ accuracy) comparison of KNN and NB algorithms for various test file size. The X-axis shows test file sizewith algorithms and Y- axis shows %.

V.Conclusion

We proposed heart disease prediction system based on symptoms. For heart disease prediction based on symptoms, we used machine learning algorithm that is Naïve bayes and CNN. We performed heart disease prediction using naïve bayes algorithm and KNN algorithm. We compare the results between KNN and Naïve bayes algorithm and the accuracy of NB algorithm is 94% which is more than KNN algorithm. We got accurate heart disease risk prediction as output, by giving the input as patients record which help us to understand the level of heart disease risk prediction. Because of this system may leads in low time consumption and minimal cost possible for disease heart disease risk prediction.

In future, we will add more diseases and predict the risk which patient suffers from specific disease.

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 Common*., 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., Issue 6, June 2018.