Disease Risk Prediction by Using Convolutional Neural Network

Sayali Ambekar
Department of Information Technology
MIT College of Engineering
Pune, India
sayaliambekar21@gmail.com

Rashmi Phalnikar
Department of Information Technology
MIT College of Engineering
Pune, India
rashmi.phalnikar@mitcoe.edu.in

***Abstract*— Data analysis plays a significant role in handling a large amount of data in the healthcare. The previous medical researches based on handling and assimilate a huge amount of hospital data instead of prediction. Due to an enormous amount of data growth in the biomedical and healthcare field the accurate analysis of medical data becomes propitious for earlier detection of disease and patient care. However, the accuracy decreases when the medical data is partially missing. To overcome the problem of missing medical data, we perform data cleaning and imputation to transform the incomplete data to complete data. We are working on heart disease prediction on the basis of the dataset with help of Naïve bayes and KNN algorithm. To extend this work, we propose the disease risk prediction using structured data. We use convolutional neural network based unimodel disease risk prediction algorithm. The prediction accuracy of CNN-UDRP algorithm reaches more than 65%. Moreover, this system answers the question related to disease which people face in their life.**

Keywords*— *Data Mining, Heart Disease Prediction, Naïve Bayes, KNN, Heart disease risk prediction, CNN-UDRP algorithm.

I. INTRODUCTION

Mostly 50% of American suffers from one or more than one chronic disease and because of that, those spent more amounts on the treatment of disease [1]. As the lifestyle is improved, then the frequency of disease is also increasing. In India, nearly 61% of death causes due to the non – communicable disease like heart disorder, cancer, and diabetes. The main reason behind the cause of diseases is environmental condition and living habits of people. To get earlier disease detection, reduced the risk of disease and diagnosis of the disease there is IOT based disease prediction used [3]. But the main drawback of IOT based disease prediction is patient needs to wear more devices and that becomes difficult and not comfortable to the patient [3]. So now a day in the healthcare field, the data mining has been a preferable technique for disease prediction, detection and diagnosis purpose. Extracting the required information from a large amount of historical data set is called data mining. Future prediction is happening on the basis past historical data [8]. There are many fields in data mining like statistics, Artificial intelligence, database technique, and machine learning [8]. Similarly in the medical field the exploration of hidden pattern done with the help of data mining [8].

Mostly the medical data contains hidden information so the decision making becomes a difficult task. The machine learning plays a significant role for finding the hidden pattern in medical data and analysis of data. There are multiple domains of the machine learning to perform the analysis of data like finance, government, transport, healthcare, and marketing etc [6]. Machine learning is the subfield of data mining which handles a large amount of well-formatted data. In the medical field, machine learning used for prediction, detection, and diagnosis of disease [6]. The Main goal of these techniques is to earlier detection of heart disease so that it may lead to earlier diagnosis of disease with good treatment on the heart disease. In past years, the medically based research done on the automatically extracted large amount of feature from structure data. In the database, there are three types of data in that consist of structured data, unstructured data, and semi-structured data. Structured data means well-formed data which consist of actual patient's record like EHR, laboratory records, some medical tests reports data etc [2]. The Motivation behind this is to handle a huge amount of heart disease data and on that the risk prediction of heart disease. The data cleaning and data imputation is necessary on medical data are not in proper format. Due to the not well-formatted data, the disease prediction is unable to perform and sometimes due to this may cause incorrect prediction of disease [1].

We already had done with the prediction of disease on the basis of symptoms. We used naïve bayes algorithm for the prediction of disease with the help of structured data. In this paper, we perform operation on medical structured data. The convolutional neural network is deep learning concept in that extracting the features automatically from the large dataset and get the proper result. For structured data, CNN-UDRP used for extracting necessary features values from the dataset and on the basis of that dataset prediction of disease carried out. The main objective of this paper is to predict the heart disease as well as heart disease risk on the basis of structured data. The second objective is handling missing values to find the accurate risk of heart disease.

The reminder of this paper is organized as follows. In section 2, we discuss the review of literature. In section 3, we design proposed system. In section 4, we present the algorithm use for disease risk prediction. In section 5, we represent the experimental setup and result obtains from this system. Finally, section 6, concludes the paper.

II. REVIEW OF LITERATURE

In paper [1], Chen proposed a new convolutional neural network based multimodal disease risk prediction algorithm by using structured and unstructured data of hospital. Authors invented disease prediction system for the various regions. Also, predict that whether a patient experiences from the high risk of cerebral infarction or low risk of cerebral infarction. The accuracy of disease prediction reaches up to the 94.8% with faster speed than Convolutional neural network based unimodal disease risk prediction algorithm.

978-1-5386-5257-2/18/$31.00 ©2018 IEEE

In paper [2], authors designed the Alzheimer disease risk prediction system with the help of EHR data of the patient. Here they used active learning context to solve a real problem suffered by the patient. The experts identify the similar health conditions between the two patients and on the basis of that patient risk correctly evaluated.

Designed cloud-based health –Cps system in [3], which manage the huge amount of biomedical data. This system performed various 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. Also, the many services related to healthcare know by this system.

In paper [4], the author proposed wearable 2.0 system in which design smart washable clothing that improves the QoE and QoS of the next-generation healthcare system. With the help of this cloth, it captured the physiological condition of the patient. And for the analysis purpose, this data is used. Discuss the issues which are facing while designed the wearable 2.0 architecture. In this, there are many applications discussed like chronic disease monitoring, elderly people care, emotion care etc.

Proposed telehealth system [5] in that author discusses how to handle a large amount of hospital data in the cloud. For this author invented new optimal big data sharing algorithm. By this algorithm, users get the optimal solution of handling biomedical data.

In paper [8], 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. And 2D/3D graph and pie charts designed for visualization purpose.

The heart disease prediction involve consist heart or blood vessels. The Diagnosis and prediction of heart diseases are most important so that can reduce the risk of disease [9]. In healthcare research, extreme works have been done on the prediction of heart diseases.

Mostly two algorithms used like CNN and Genetic algorithm for the prediction of heart disease. And also here major factors are considering age, family history, diabetes, hypertension, cholesterol, smoking, alcohol intake, obesity or physical inactivity etc [10].

In paper [11], for the heart disease prediction system design, authors were using machine learning algorithms like Naive Bayes, Neural network and Decision tree algorithms.

III. PROPOSED SYSTEM

The Main idea of this paper is to predict whether the patient suffers from heart disease or not. And also predicting the risk of heart disease that is patient it is at high risk or low risk. The user enters the appropriate input values from his/her health report. After this, the historical dataset is uploaded. The mostly medical dataset may contain missing values due to this accurate prediction becomes difficult. So for this missing data imputation and data cleaning step is necessary. After this data Imputation, we have to transform the missing data into structured data with the help of a data cleaning and data imputation process. Then the naïve bayes and KNN algorithm is implemented on the input values and on the bases of this heart disease is predicted. The two algorithms consider here first one is naïve bayes and KNN for classification. But the classifier which gives highest accuracy value that classifier value gave as input to the CNN-UDRP algorithm for risk prediction. Between this two classifiers naïve bayes classifier performance is better so the CNN-UDRP gets the input from Naïve bayes classifier. By using CNN- UDRP algorithm, we can predict whether the patient suffers from high or low risk. The Convolutional neural network algorithm is used for feature extraction purpose. And on that features, the softmax classifier is apply to get the classification of heart disease risk.

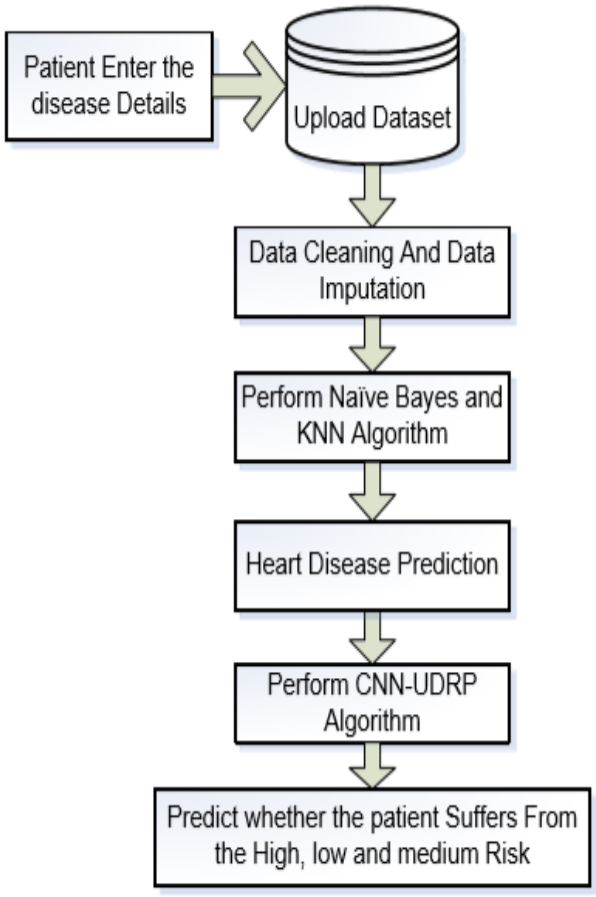

Fig 1: Proposed System Architecture

A. Dataset

The heart disease prediction and for heart risk prediction we are using dataset from UCI Repository. The dataset consist of 12 attributes which are age, sex, Chest Pain Type, Blood Pressure, Cholesterol, Resting results of Electrocardiography, number of major vessels that colored by flourosopy, fast Blood Sugar, ST segments that induced by exercise relative to rest , Peak exercise ST segments, Defect values, Exercise induced angina etc. On the basis of these parameters the actual heart disease prediction carried out. And the risk of heart disease is carried out on the result of KNN and Naïve byes algorithm but here mostly we are considering Naïve Bayes algorithm result for further risk prediction in percentage and displaying high risk, low risk or medium risk of the patient suffer from high disease problem.

Table1: Attributes of heart disease and heart disease risk prediction.

No.	Attributes Name	Description
1.	Age	Age of the patient in a year
2.	Sex	Gender of the patient
3.	Cp	Chest Pain Type
4.	Trestbps	Blood Pressure
5.	Chol	Cholesterol
6.	Restecg	Resting results of Electrocardiography
7.	Ca	Number of major vessels that colored by flourosopy
8.	Fbs	Fast Blood Sugar
9.	Oldpeak	ST segments that induced by exercise relative to rest
10.	Slope	Peak exercise ST segments
11.	Thal	Defect values
12.	Exang	Exercise induced angina

B. Data cleaning and data imputation

In dataset, it consists of unstructured data means data which is not in well-formed data. Mostly medical data is not in proper format. For the missing data, the data cleaning and data imputation is necessary. The unwanted data and noisy data must remove from dataset so that we get structured data.

C. Disease Risk prediction

The main motive of this paper is to predict the patient is among the heart disease high or low risk according to the heart disease dataset. We design the risk prediction model for heart disease as the deep learning algorithm for that input as we taking as age, sex, thal, Blood pressure etc. The output value is h, which shows that whether the patient is among the high risk or low risk as h= {h1, h2}, where as h1 predict the high risk of heart disease whereas h2 predict the low risk of heart disease.

D. EVALUATION METHOD

For the experimental result, we have notations as follows:
TP: True Positive (number of instance which correctly predicted),
FP: False positive (number of instance which incorrectly predicted),
TN: True negative (correctly predicted the number of instances as not required),
FN: false negative (incorrectly predicted the number of instances as not required),

On the basis of this parameter, we can calculate four measurements

1. ACCURACY = TP+TN÷TP+FP+TN+FN
2. PRECISION = TP ÷TP+FP
3. Recall= TP÷TP+FN
4. F1-Measure = 2 × Precision × Recall ÷ Precision +Recall.

To get more benefits of system user can also post their query onto the system so that patient knows about any kind of answer related to disease from other doctor or specialist. The advantage of this system is to get know about the risk associated with the heart disease.

IV. ALGORITHM

For the risk prediction of disease, we are using deep learning algorithm. The convolutional neural network is deep learning algorithm; this algorithm extract text features automatically. Here we are predicting risk associated with that disease by the CNN-UDRP algorithm. Before that it performs the heart disease prediction by using Naïve bayes and KNN algorithm.

A.Naïve Byaes

Naïve bayes classifier based on probabilistic model and depends on the bayes theorem. In the supervised learning, the naïve bayes classifier work. The particular features which is describe in a class that are not related to the another features.

P(c/y) = P(y/c)* P(C)/ P(y)
P(c/y) = posterior probability,
P(c) = prior probability of class,
P(y/c) = likelihood probability of the class,
P(y) = prior probability of predictor.

On the bases of this algorithm, the classification is carried out.

B.KNN Algorithm

KNN is a classifier which stored all the values of the variable that is records and on the bases of that records, the unknown value of variable is classify. The unknown value is classified among the similarity of the variable. KNN is a non parametric classification method. The KNN is divided into two types the first one is structure less NN technique and second one is structure based NN technique. The structured based NN in that data is classified into training and testing data.

C. CNN-UDRP Algorithm

For the prediction of risk associated with heart disease, we have to perform five steps of algorithm. In first step the dataset is converted into the vector form. Then word embedding carried out which adopt zero values to fill the data. The output of word embedding is convolutional layer. This convolutional layer taken as input to pooling layer and we perform max pooling operation on convolutional layer. Pooling layer is connected with the full connected neural network. Lastly, the full connection layer connected to the classifier that is softmax classifier.

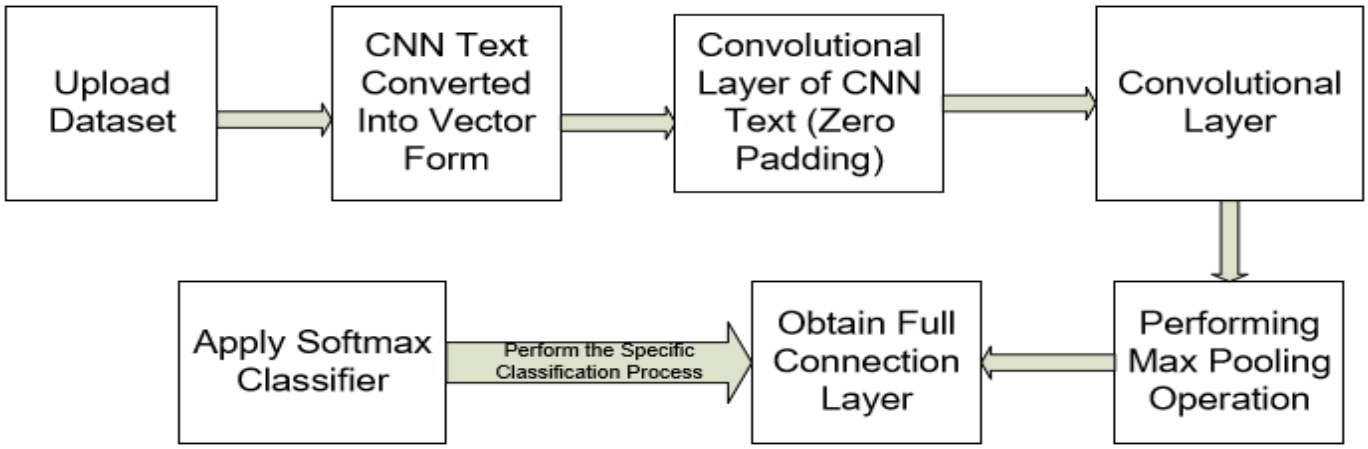

Fig 2: CNN-UDRP Algorithm Steps

V. Experimental Stepup

We performed heart disease prediction and heart disease risk prediction on dataset. In the dataset contain training and test dataset. We divide the dataset into training which contains near about 800 entries and 200 entries of test data. The accuracy of heart disease prediction of naïve bayes and KNN algorithm is compare. The naïve bayes algorithm accuracy is near about 82% which is more than KNN algorithm.

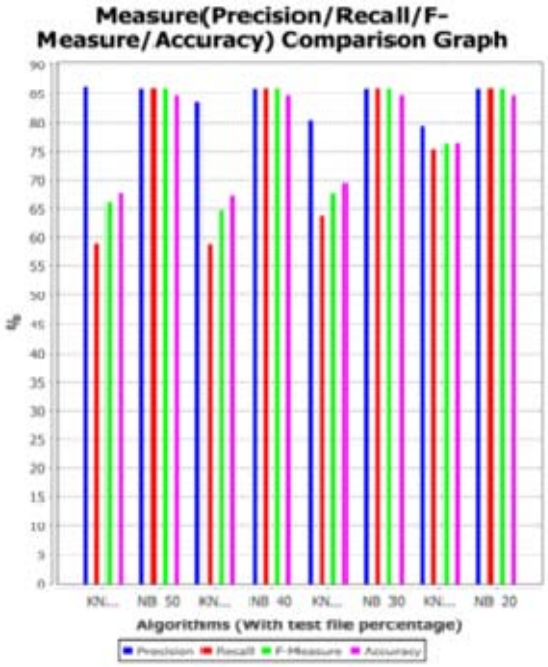

Fig 3: The accuracy comparison between KNN and Naïve Bayes Algorithm

A. Parametes Comparsion

As we changes the training data set and testing dataset values the accuracy of the KNN and Naïve bayes algorithm is changed. Here the accuracy of the Naïve bayes algorithm is always more as compare to the KNN algorithm. The other parameters like recall, precision is also varies according to the training and test data is changes.

B. Time Comparsion

Also the time is an important factor for any algorithm, so how much time needed for the execution of algorithm is significant factor. Here two algorithm used for heart disease prediction, so from these two algorithm that is KNN and Naïve Bayes we calculate time according to the varying in training and testing dataset.

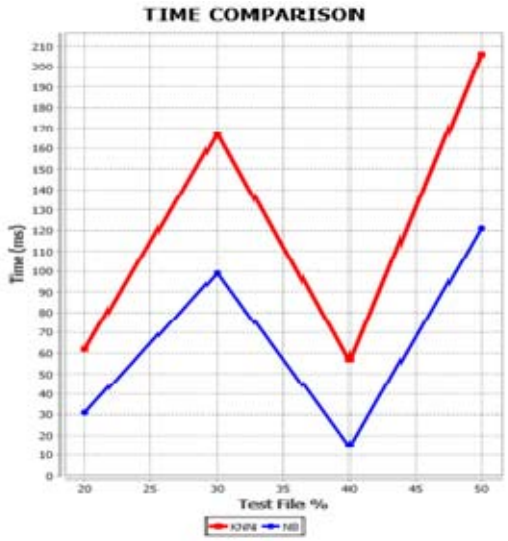

Fig 4: The execution time for KNN and NB algorithm

Figure 4 shows that On X- axis shows the test data which is used for execution of the algorithm. And on Y-axis shows the time required to run the algorithm. It is measure in millisecond. Here in graph shows that the time required running the algorithms at different testing dataset. The blue line shows that time required running the Naïve bayes algorithm and the red line shows the time required to run the KNN algorithm. By this graph we got the result as follows:

Table 2: Time required for execution of two algorithms

NO.	Training Dataset	Testing Dataset	Naïve Bayes	KNN algorithm
1.	80%	20%	30ms	60ms
2.	70%	30%	90ms	175ms
3.	60%	40%	20ms	55ms
4.	50%	50%	120ms	208ms

As in table and graph shows that as the KNN required for time to execute as compare to the Naïve bayes algorithm. So we mostly prefer Naïve bayes algorithm result for further heart disease risk prediction.

C. CNN-UDRP Performance

For Heart disease risk prediction, we used CNN-UDRP algorithm, the main CNN algorithm perform number of iteration for obtaining accurate result. Here we considered 500 iteration and input layer contain 10 input factors for heart disease prediction purpose. The only one output is display which contains percentage that how many percentages the patient actually suffer from heart disease risk.

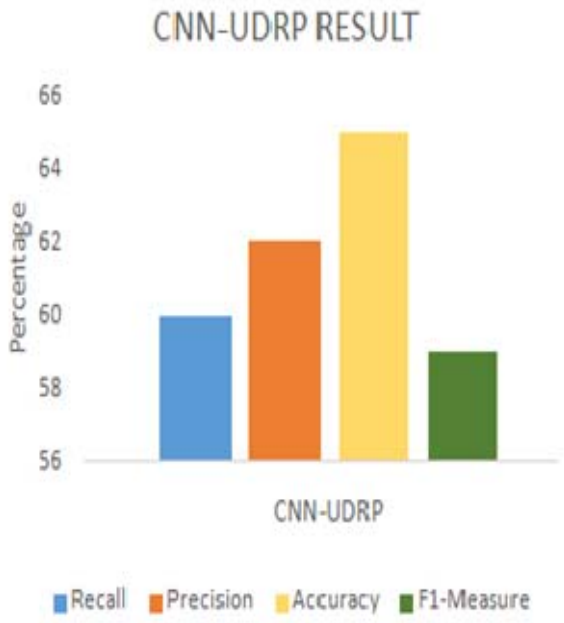

Fig 5: Performance of CNN-UDRP algorithm

In Figure 5: Shows the various parameters like precision, recall, F1-measure and accuracy. The accuracy of heart disease risk prediction is near about 65%. The recall and precision values are near to 60%, 62%. So the performance of CNN-UDRP is better for heart disease risk prediction. The result is display to patient in the form of percentage and high, low and medium risk result display to the patient.

VI. Conclusion

In this paper, we experiment the CNN-UDRP algorithm using structured data for disease risk prediction. 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 82% which is more than KNN algorithm. We got near about 65% of accuracy of disease risk prediction with the help of structured. We got accurate disease risk prediction as output, by giving the input as patients record which help us to understand the level of disease risk prediction. The risk is predicted as low, high and medium risk of heart disease. Because of this system may leads in low time consumption and minimal cost possible for 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," DataMiningKnowl.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.23-25.

[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), 2013, 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.