Machine Learning in Smart HealthCare Industry

Lithu Mathew[1], Kala O S[2]

[1,2]*Assistant Professor, Department of Computer Science and Engineering, Al Azhar College of Engineering and Technology, Thodupuzha*

Abstract - **Chronic disease prediction plays an important role in healthcare informatics. It is crucial to diagnose the chronic disease at an early stage. In the field of healthcare communities, the accurate analysis and prediction plays the major role to find out the risk of the disease in the patient. However, the analysis of accuracy is reduced, and it leads to less accuracy of prediction when the quality of data is incomplete and the poor condition of the medical data. We seek machine learning techniques for effective prediction of chronic disease. We propose to use convolution neural network algorithm for structured and unstructured data. The accuracy obtained using CNN model reaches 94.8% and none of the existing algorithm focuses on structured and unstructured data in the area of medical data analytics.**

Index Terms – **Convolution neural network, structured data, Data imputation.**

I.INTRODUCTION

Machine learning is programming computers to optimize a performance using sample data or past data. Machine learning is the study of computer systems that learn from a data set. Machine Learning technology provides a good platform in the medical field, so that a healthcare issue can be solved efficiently. Disease prediction using patient treatment history and health information by applying data mining techniques are ongoing struggle for the last few years.

The healthcare and medical field are more in need of datamining today. Disease and health related problem like malaria, dengue, impetigo, and cancer etc.it is reported that 60% of Indians suffer from at least one chronic disease. As the living standards raises, the effect of these diseases also increases. The India is not the only country where large sums are spent treating chronic diseases.in China, USA, most people's die because of chronic diseases. Chronic diseases are the main cause of death and according to a report on nutrition and chronic diseases in 2015,86% of death are caused by chronic diseases. Clearly, it is essential that early diagnosis and treatment are essential, not just to save costs, but also save human life and improve quality of life.

Big data analytics is the process of examination of huge data sets which consists of variety of data types. Concept of the big data is not a new concept which is constantly changing.

The healthcare data is spread among the multiple medical systems, healthcare sector and hospitals with the benefits of a big data in which more attention is paid to the disease prediction. We can say that Machine learning is an application of artificial intelligence (AI). This provides the system, ability to automatically learn and improve from experience without being explicitly programmed.

With the advancement of machine learning technology, more attention has been paid to disease prediction from the perspective of big data analysis, various research have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification [1], [2], rather than the previously selected characteristics. Most of the existing work considered the structured data. For unstructured data, using convolutional neural network (CNN) to extract text characteristics automatically has already attracted a big attention and also achieved good results [3], [7]. However, none of previous work handles the medical text data by CNN. Furthermore, the disease in different regions varies, primarily because of the diverse climate and living habits in the region. In the risk classification based on big data analysis, the following challenges remain: How should the missing data be addressed? and how should the main chronic diseases in a certain region and the main characteristics of the disease in the region be determined? How can big data analysis and machine learning technology be used to predict the disease and create a better model? As a solution we combine the structured and unstructured data in healthcare field to predict and assess the risk of disease.

For structured and unstructured data, we propose a CNN- based multimodal disease risk prediction (CNN-MDRP) algorithm. The risk model of disease is obtained by the combination of structured and unstructured features. Through the experiment, we can conclude that the performance of CNN-MDPR is better than other existing methods.

II. MODEL DESCRIPTION AND EVALUATION METHOD

A. data set

The dataset used in this study contains real-life hospital data, and the data are stored in the data center. The medical tests that we have considered in this study are hospital data and same are stored in our database. A security access method is used to protect the patient's privacy and security. a large volume of datasets of patient can be given by a hospital which can be processed in the information center to preserve the patient security and privacy. The data provided by the hospital include electronic health records, medical image data and gene data. We use a four-year data set from 2015 to2019. The inpatient department data is mainly focused on structured and unstructured text data. The structured data includes laboratory data and the patient's basic information such as the patient's age, gender etc. While the unstructured data includes the patient's description of his/her illness, the doctor's interrogation records etc. They are classified into two categories structured data and unstructured text data.

B. Chronic Disease Prediction

Here, in this modal we are mainly focused on prediction of chronic diseases. The aim of the designed model is to predict whether a patient is amongst the cerebral infarction high-risk population according to their medical history. We regard the risk prediction model for the disease as the supervised learning methods of machine learning. The value of input is the attribute value of the patient, X = (x1, x2, · · ·, xn). This vector contains the patient's personal information such as age, gender, the prevalence of symptoms, and living habits and other structured data and unstructured data. The output value(C) indicates whether the patient is amongst the cerebral infarction high-risk population. C = {C0, C1}, where, C0 represents the patient is at high-risk of cerebral infarction, C1 represents the patient is at low-risk of chronic disease.

• Structured data (S-data): To predict whether the patient is at high-risk of chronic disease.

• Text data (T-data): use the patient's unstructured text data is used to predict whether the patient is at high-risk of chronic disease.

• Structured and text data (S&T-data): The S-data and T-data use both the structured data and unstructured text data to predict whether the patient is at high-risk of chronic disease.

In the experiment, we will introduce machine learning and deep learning algorithms used to predict chronic diseases. For S-data, we use conventional machine learning algorithms such as Naive Bayesian (NB), K-nearest Neighbor (KNN), and Decision Tree (DT) algorithm [5], [8] to predict the risk of chronic diseases. For T-data, a CNN- based unimodal disease risk prediction (CNN-UDRP) algorithm to predict the risk of chronic disease. In the remaining experiments, let CNN-UDRP(T-data) denote the CNN-UDRP algorithm used for T-data. And S&T data, we predict the risk of chronic disease by the use of CNN-MDRP algorithm, which is denoted by CNN-MDRP(S&T-data) for the sake of simplicity.

C. Performance Evaluation

For the performance evaluation, we denote TP, FP, TN and FN as true positive (the number of instances correctly predicted as required), false positive (the number of instances incorrectly predicted as required), true negative(the number of instances correctly predicted as not required) and false negative (the number of instances incorrectly predicted as not required), respectively. The four measurements obtained are: accuracy, precision, recall and F1-measure as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Where the F1-Measure is the weighted harmonic mean of the precision and recall and represents the overall performance. In addition to the aforementioned evaluation criteria, we use receiver operating characteristic (ROC) curve and the area under curve (AUC) to evaluate the classifier. The Receiver Operator Characteristic curve shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR), where the TPR and FPR are defined as follows:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

If the ROC curve is near to the upper left corner of the graph, the model is better. The AUC is the area under the curve. When the area is closer to one, then the model is better. In medical data, we given more attention to the recall rather than accuracy. The higher recall rate, the lower the probability that a patient who will have the risk of disease is predicted to have no disease risk.

III. PROPOSED SYSTEM

In the proposed method, we are introducing the data imputation, CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm and CNN-based unimodal disease risk prediction (CNN-MDRP) algorithm.

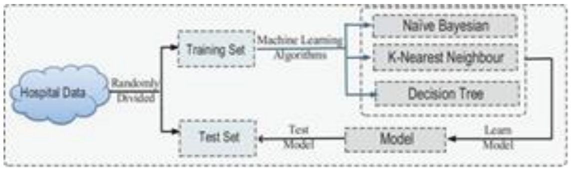
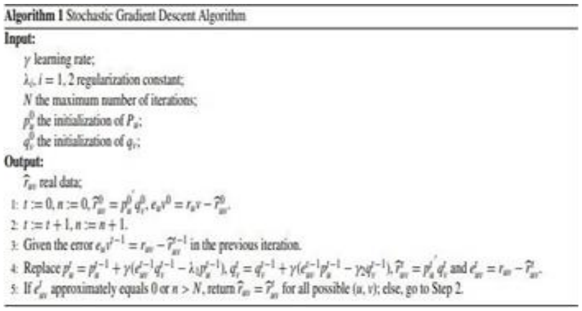


Figure 1: Proposed chronic disease prediction model

A. Data Imputation

For patient's examination data, there is a huge number of missing data due to human error. Thus, we need to fill out the structured data. Before data imputation, we need to identify uncertain or incomplete medical data and then modify or delete them to improve the data quality. Data pre- processing methods are used for data integration. We can integrate the medical data to guarantee data atomicity and integrated the height and weight to obtain body mass index (BMI). For data imputation, we use the latent factor model [6] which is presented to explain the observable variables in terms of the latent variables.



```
Algorithm 1 Stochastic Gradient Descent Algorithm
Input:
    $\gamma$ learning rate;
    $\lambda_i, i = 1, 2$ regularization constant;
    $N$ the maximum number of iterations;
    $p_u^0$ the initialization of $P_u$;
    $q_v^0$ the initialization of $q_v$;
Output:
    $\widehat{r}_{uv}$ real data;
1: $t := 0, n := 0, \widehat{r}_{uv}^0 = p_u^{0'} q_v^0, e_u v^0 = r_u v - \widehat{r}_{uv}^0$.
2: $t := t + 1, n := n + 1$.
3: Given the error $e_u v^{t-1} = r_{uv} - \widehat{r}_{uv}^{t-1}$ in the previous iteration.
4: Replace $p_u^t = p_u^{t-1} + \gamma(e_{uv}^{t-1} q_v^{t-1} - \lambda_1 p_u^{t-1})$, $q_v^t = q_v^{t-1} + \gamma(e_{uv}^{t-1} p_u^{t-1} - \gamma_2 q_v^{t-1})$, $\widehat{r}_{uv}^t = p_u^{t'} q_v^t$ and $e_{uv}^t = r_{uv} - \widehat{r}_{uv}^t$.
5: If $e_{uv}^t$ approximately equals 0 or $n > N$, return $\widehat{r}_{uv} = \widehat{r}_{uv}^t$ for all possible $(u, v)$; else, go to Step 2.
```

We can solve it by the use of the stochastic gradient descent method. We can get the specific solution as shown in Algorithm 1, which can fill missing data.

B. CNN-UDRP algorithm

In the processing of medical data, we use CNN-based uni modal disease risk prediction (CNN-UDRP) algorithm which can be divided into the following five steps.

1) Text Data Representation

Usually, each word in the medical text is represented in the form of vector. In this experiment, each word will be represented as a R^d -dimensional vector, where d = 50. Thus, a text including n words can be represented as T = (t1, t2, · · ·, tn), $T \in R^{d\times n}$.

2) Convolution Layer of Text CNN

The first layer is the convolution layer. The Layer retains the relationship between pixels of image. It does this by learning its properties. This method is performed by dividing the image into smaller box of pixels.

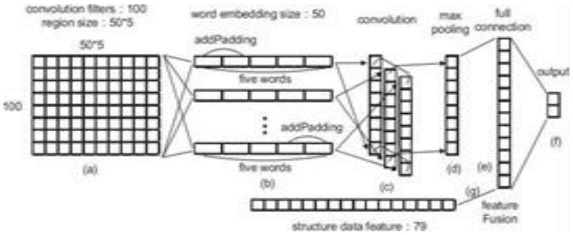


Fig 2: The model's Basic Framework

3) Pool Layer of Text CNN

Pooling layer is a building block of CNN. It reduces the number of parameters and computation used in the network.

The output of convolution layer is taken as the input of pooling layer, we use the max pooling (1-max pooling) operation. It's shown in Fig 1. Select the max value of the n elements of each row in feature graph matrix

$h^1 : h^2_j = \max_{1\leq i\leq n} h^1_{i,j}$, j = 1, 2, · · · , 100

After max pooling, we obtain 100×1 features h^2 . The main reason of choosing max pooling operation is that the role of every word in the text is not completely equal; by maximum pooling we can choose the elements which play a vital role in the text.

4)Full Connection Layer of Text CNN

Pooling layer relates to a fully connected neural network as shown in Fig. 1. The specific calculation process is that:

$h^3 = W^3h^2 + b^3$

where h^3 is the value of the full connection layer, W^3 and b^3 is the corresponding weights and deviation.

5) CNN Classifier

The full connection layer links to a classifier, for the classifier, we need to choose a softmax classifier, as shown in Fig. 1(f).

C. CNN-Based Multimodal Disease Risk Prediction (CNN-MDRP) Algorithm

We have the information that CNN-UDRP only uses the text data to predict whether the patient is at high risk of disease. Unstructured data is important to fill the gaps in the structured data. For the structured and unstructured text data, a CNN-MDRP algorithm is designed based on CNN-UDRP as shown in Fig. 1. The only algorithm used to predict whether the patient is at high risk of chronic disease is CNN-MDRP. The processing of text data is similar with CNN- UDRP, which can extract 100 characteristics about text data set. For structure data, we extract 79 features. Then, we conduct the feature level fusion by using these features in the S-data and 100 features in T-data. For full connection layer, similar computation methods are used with CNN- UDRP algorithm since the variation of features number, the corresponding weight matrix and bias change to W^3 new, b^3 new, respectively.

We also utilize softmax classifier. The specific training process is divided into two parts. Training word embedding and training parameters of CNN-MDRP. In CNN-MDRP algorithm, the specific training parameters are w^1,w^3,b^1,b^3 we use $_{new}$ stochastic $_{new}$ gradient method to train parameters, and finally reach the high risk assessment of whether the patient suffer from chronic disease.

IV. RESULTS AND DISCUSSIONS

Fig 3 shows the accuracy, Recall of T-data and S&T –data using CNN-UDRP and CNN-MDRP algorithm .here we set the same CNN iteration which are 100 and extract the same 100text features. We can see the running time of CNN – UDRP (T-data) and CNN-MDRP (S&T data) are basically the same from the figure. ie, the number of CNN-MDRP features increase after adding structured data. It does not make significant changes in time.

V. PERFORMANCE ANALYSIS

For the performance evaluation of this model, we denote TP, FP, FN and TN. Then, we can find four measurements: accuracy, precision, recall. In addition to the evaluation criteria, we use receiver operating characteristics (ROC) curve and the area under curve (AUC) to evaluate the pros and cons of the classifier. The metrics provided below given us information on the quality of the results that we get in this study.

Disease	Accuracy	Correctly classified instance	Incorrectly classified instance
Cancer	81.8182	108	24
Heart	99.8243	568	1
Diabetes	78.9272	206	55
Kidney	77.2121	309	91

Table 1: Result on accuracy with correctly classified and incorrectly classified instance.

Disease	Precision	Recall	f-Measure
Cancer	1.0000	.333	.500
Heart	1.0000	.995	.998
Diabetes	0.812	.899	.853
Kidney	0.857	.666	.749

Table 2: Result of Precision, Recall, F-Measure

Traditional machine learning algorithms are used for S-data. i.e., NB, KNN and Decision Tree algorithms are used to predict the risk of chronic diseases.

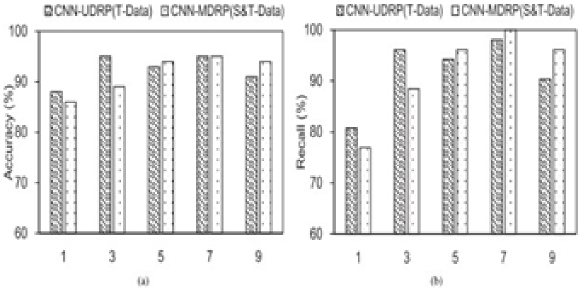


Fig 3: (a) the corresponding accuracy of the CNN-UDRP and CNN- MDRP algorithms (b) the corresponding recall of the CNN-UDRP and CNN-MDRP algorithm

It requires calculating the probability of feature attributes. In this method, we use conditional probability formula to estimate discrete feature attributes and Gaussian distribution to estimate continuous feature attributes. The KNN classification is given a training data set, and the nearest k instance in the training data set is found. For KNN, it is required to determine the value of k and the measurement of distance. In the method, the data is normalized at first. Then we use the Euclidean distance to measure the distance. The model is best when k=10. Thus, we choose k = 10. To determine the best classifier and improve the accuracy of the model, the 10-fold cross-validation method is used for the training set. The data from the test set are not used in the training phase. In summary we are used different algorithm to predict chronic disease, namely CNN, naïve Bayes, etc. It is observed that the model accuracy is highest for the model which is designed using CNN .for S-data, the NB classification is the best in experiment. However, it is also observed that we cannot predict whether the patient is in a higher risk of chronic disease based on the patient's age, gender and other structured data. In structured and text data. We assured the accuracy, precision, recall, F1-measure and ROC curve under CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms. Thus, we concluded that the accuracy of CNN-UDRP (T-data) and CNN-MDRP (S&T- data) algorithms have little difference but the recall of CNN-MDRP (S&T-data) algorithm is higher and its convergence speed is faster. In summary, the performance of CNN- MDRP (S&T-data) is better than CNN- UDRP (T-data). In conclusion, for disease risk modelling, the accuracy of risk prediction depends on the diversity feature of the hospital data. For some simple disease, e.g., hyperlipidemia, only a few features of structured data can get a good description of the disease, But for a complex disease, such as chronic disease mentioned in the paper, only using features of structured data is not a good way to describe the disease. Therefore, we are using 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 94.80%, so as to better evaluate the risk of cerebral infarction disease.

IV.CONCLUSION

Using structured and unstructured data from hospital, this paper proposes a convolution neural network based multimodal disease risk prediction (CNN-MDRP) algorithm. For existing calculation machine learning technique such as naive-Bayes was utilized. CNN-UDRP just uses organized information however in CNN-MDRP concentrated on both organized and unstructured information. None of the existing work focused on both data types. The prediction accuracy of the proposed algorithm reaches 94.8%. Using the surveyed chronic disease prediction model, we can predict before it reaches a point of no return and hence appropriate diagnoses can be performed on patients to help them as far as possible.

REFERENCES

[1] 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

[2] W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[3] J. Wan et al., ‘‘A manufacturing big data solution for active preventive maintenance,’’ IEEE Trans. Ind. Informat., to be published, doi:10.1109/TII.2017. 2670505.

[4] B. Smith, “An approach to graphs of linear forms (Unpublished work style),” unpublished.

[5] K. Hwang and M. Chen, Big Data Analytics for Cloud/IoT and Cognitive Computing. Hoboken, NJ, USA: Wiley,2017.

[6] J.C.Ho,C.H.Lee,andJ.Ghosh,‘‘Septicshockpredictionforpatientswith missing data,’’ ACM Trans. Manage. Inf. Syst., vol. 5, no. 1, p. 1,2014W. Yin and H. Schutze, ‘‘Convolutional neural network for paraphrase identification,’’ in Proc. HLT-NAACL, 2015, pp.901–911.

[7] H. Chen, R. H. Chiang, and V. C. Storey, ‘‘Business intelligence and analytics: From big data to big impact,’’ MIS Quart., vol. 36, no. 4, pp. 1165–1188, 2012.

[8] Hamet P., Tremblay J. Artificial intelligence in medicine. Metabolism. 2017; 69:S36–S40. doi: 10.1016/j.metabol.2017.01.011.[PubMed] [CrossRef] [Google Scholar]