Disease Prediction by Machine Learning over Big Data from Health Care Communities

A SEMINAR REPORT

Submitted to

Visvesvaraya Technological University

BELAGAVI - 590 018

by

NIPUN HEGDE **USN: 4SU17CS049**

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Engineering

Department of Computer Science & Engineering

SDM INSTITUTE OF TECHNOLOGYUJIRE - 574 240

2020-2021

SDM INSTITUTE OF TECHNOLOGY

(Affiliated to Visvesvaraya Technological University, Belagavi)

UJIRE – 574 240

Department of Computer Science & Engineering

CERTIFICATE

Certified that the Technical Seminar Report titled **'Disease Prediction by Machine learning over Big Data from health care communties'** is carried out by **Mr.NIPUN HEGDE** USN: **4SU17CS049** a bond-fide student of SDM Institute of Technology, Ujire in partial fulfillment for the award of thedegree of **Bachelor of Engineering** in Computer Science & Engineering of Visvesvaraya Technological University, Belagavi during the year 2020- 2021. It is certified that all the corrections/ suggestions indicated for InternalAssessment have been incorporated in the report deposited in the departmental library. The report has been approved as it satisfies the academic requirements in respect of Technical Seminar prescribed for the said Degree.

Examiners **Signature with Date**

Guide/Coordinator/HOD

1.

2.

Acknowledgement

I express my deepest gratitude to my guide Prof. Chintesh R, Assistant Professor, Department of Computer Science & Engineering, SDM Institute of Technology for his valuable guidance and encouragement while doing this technical seminar.

We are indebted to Dr. Thyagaraju G S,Head of the Department, Dr. Ashok Kumar T,Principal, Seminar coordinators Mr. Amith K S for their advice and suggestions at various stages of the work.

I am also grateful to the co-operation and help rendered by the teaching and non- teaching staff of the department.

Nipun Hegde

USN: 4SU17CS049

Abstract

With huge information headway in biomedical and healthcare communities, appropriate examination of therapeutic information helps early sickness identification, tolerant consideration and network administrations. Prediction accuracy is diminished when the nature of medicinal information is inadequate. At that point the various areas appear, one of kind qualities of certain local infections, which may debilitate the expectation of illness episodes. In this paper, machine learning method is applied for viable forecast of interminable disease in the history of predicting diseases. The main intension is to have different prediction models over genuine medical clinic information. To conquer the trouble of deficient information, a latent factor model is used to regenerate the irrecoverable data. Here, experiment on a territorial chronic infection of cerebral localized necrosis is done. CNN-MDRP (convolutional neural system based multimodal infection chance prediction) algorithm is explained utilizing organized and unstructured information from medical clinic. Apparently, none of the current work establishes on the two information types in the zone of therapeutic enormous information investigation. Contrasted with numerous prediction algorithm, the precision accuracy of the proposed method arrives with a combination speed which is quicker than that of the CNN-UDRP(convolutional neural network based unimodal disease risk prediction) Convolutional Neural Network Based Multimodal Disease Prediction(CNN-MDRP) algorithm is overcome the drawbacks of CNN-UDRP algorithm only focus work on a structured data but CNN-MDRP algorithm uses both structured and unstructured data from the hospital. None of the existing work focused on both datatypes in the area of medical big data analysis .CNN-MDRP algorithm prediction is more accurate than compared to the previous prediction algorithm.

Table of Contents

List of Figures

CHAPTER 1

INTRODUCTION

With the improvement of living standards, the incidence of chronic disease is increasing. According to the National Health Profile 2019, India's per capital public expenditure on health in nominal terms is Rs 1,657 (2018-19), which is much lower than countries like Sri Lanka where it is three times more and Indonesia where it is twice more than India.The Centers for Disease Control and Prevention (CDC) estimates that 90 percent of national healthcare spending goes toward chronic disease management and mental healthcare, which means that strong mental health .With the advance of big data analytic equipment, more devotion has been paid to disease expectation from the perception of big data inquiry, various explores have been conducted by choosing the features mechanically from a large number of data to improve the truth of menace classification rather than the formerly selected physiognomies.

Concept of the big data is not a new concept which is constantly changing. Big data is nothing but the collection of data. There are three important v's of data such as velocity, volume and variety. Healthcare is a best example of adapting these three v's of data. The healthcare data is spread among the multiple medical systems, healthcare sectors, and government hospitals with the benefits of a big data in which more attention is paid to the Disease Prediction. Numbers of researches have been conducted to selecting the characteristics of a disease prediction from a large volume of a data. Most of the existing works were based on a structured data. For the unstructured data, one can use a convolutional neural network which is made up of a neurons, each neurons receives some inputs and performs operations and the whole network expresses a single differentiable score functions. The accuracy of a disease prediction can be reduced because there is a more difference in a various regional disease because of climate and living habits of the peoples in their particular regions.We combine both the structured and unstructured data.to accurately predict the disease overcome the problem of a missing and incomplete data .we can use a latent factor model. In the previous work only structured data can be used but for the accurate results we can use the unstructured data. We can select characteristic automatically using CNN algorithm. WE can purpose a CNN-MDRP algorithm for both the data types. We can use machine learning algorithm for more accurate results.

CHAPTER 2

LITERATURE REVIEW

2.1 GENERAL INTRODUCTION

Literature Survey is an important Activity, which is done while gathering information about a particular topic. It will help us to get required information or ideas to do work. Recent research show possibility of exploiting video-based side channels to steal the smart device user's sensitive information.

2.2 LITERATURE SURVEY

Hen-Ying Hung, Wei-Chen Chen,Machine Learning Algorithms for Stroke Prediction in a Large-Scale Population-Based Electronic Medical Claims Database, 2017 IEEE. the researchers present how artificial intelli-gence applied to medical field for the efficient diagnosis.

For that purpose they use a k nearest neighbours algorithm and they check the accuracy of the algorithm with the help of UCI machine learning repository datasets. They had to generate patients input and test data for diagnosis. They use a real patient data. They add a additional training sets allow more medical conditions to be classified with the minimal no of changes to the algorithm. [2]

In this paper, they applying a machine learning techniques by using EMC'S from outpatients department and the algo-rithm are based on a DNN AND DBDT, It can be achieve a high UAR for predicting the future stroke prediction. It provides a several advantages like high accuracy, fastest prediction, and consistency of results. DNN algorithm also requires a lesser amount of data. DNN algorithm can achieves a optimal results by using a lesser amount of a patient data than compared to the GDBT algorithm. [3]

Richard Osuala and Ognjen Arandjelovi c University of St Andrews, They used a PCA to reduce the no of attributes, after reducing the size of the datasets; SVM can outperform a Neavi Bayes and Decision tree. SVM can also be used for prediction of hearts disease. The main goal of this paper is to predict the diabetics disease. Using a WEKA data mining tools. Data mining is very useful techniques used by health care sector for classification of disease. The aim of this paper is to study supervised machine learning algorithm to predict the heart disease. [7]

In this paper the data mining and the big data in the healthcare sector is introduced. Machine learning algorithm has been used to study the healthcare data. The continuous increase of data in a healthcare. Several

countries are spending a lot of resources, scientist leads to fix the problem of storage and organization of data the data mining will help exploitation complexity of the data and find out the new result this paper is based on the use of data mining and big data in the healthcare sector. [9]

Chen, Y. Ma, J. Song, C. Lai, B. Hu, Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring, In [9], the authors proposed a CNNMDRP (convolutional neural network based multimodal disease risk prediction) which overcomes the drawbacks of CNN-UDRP (convolutional neural network based unimodal disease risk prediction). This algorithm uses both the structured and unstructured data of a hospital compared with other existing algorithm which can work on either the structured or unstructured data. Authors have explained that the proposed algorithm produced the accuracy of 94.8. In this paper, the researchers presented how artificial intelligence applied to medical field for the efficient diagnosis. Also, to fulfill this need, authors used a k nearest neighbour's algorithm and verified the accuracy of the algorithm with the help of UCI machine learning repository datasets. In addition to this,it is needed to generate patients input along with test data for diagnosis. Authors have considered a real patient data for which additional training sets were added which allow more medical conditions to be classified with the minimal no of changes in the algorithm.

In this paper [12], authors have applied a machine learning techniques by using EMC'S from outpatients department and the algorithm were based on a DNN AND DBDT which can achieve a high UAR for predicting the future stroke problem. This technique provides a several advantages such as high accuracy, fastest prediction, and consistency of results.

In [5], the data mining along with big data in the healthcare sector was elaborated for which Machine learning algorithm has been used to examine the healthcare data. The continuous increase of data in a healthcare sector, several countries is spending a lot of resources, scientist help to cure the issues of space and establishment of data. Also, data mining will help exploitation complexity of the data and find out the new result which is based on the use of data mining and big data in the healthcare sector.

CHAPTER 3

PROBLEM STATEMENT

The major challenge in any disesase prediction is its right detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of disease in human. Early detection of cardiac diseases or any such diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

At the same time treatment of the said disease is quite high and not affordable by most of the patients particularly in India. Almost all the hospitals use some hospital management system to manage healthcare in patients. Unfortunately most of the systems rarely use the huge clinical data where vital information is hidden. As these systems create huge amount of data in varied forms but this data is seldom visited and remain untapped. So, in this direction lots of efforts are required to make intelligent decisions. The diagnosis of this disease using different features or symptoms is a complex activity. The early prognosis of diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. When the quality of medical data is incomplete the exactness of study is reduced. Moreover, different regions exhibit unique appearances of certain regional diseases, which may results in weakening the prediction of disease outbreaks.At that point the various areas appear, one of kind qualities of certain local infections, which may debilitate the expectation of illness .Moreover the accuracy of an analysis can be reduced due to an various reason like incomplete medical data, some regional disease characteristics which can be outbreaks the prediction. .

CHAPTER 4

PROPOSED METHODOLOGY

4.1 Methodology

To solve these problems, we combine the Structured and Unstructured data in healthcare field to assess the risk of disease. First, we used latent factor model to reconstruct the missing data from the medical records collected from a hospital Second, by using statistical knowledge, we could determine the major chronic diseases in the region. Third, to handle structured data, we extract useful features. For unstructured text data, we select the features automatically using CNN algorithm.

Data category	Item	Description
Structured data	Demographics of the patient	Patient's gender, age, height, weight, etc.
	Living habits	Whether the patient smokes, has a genetic history, etc.
	Examination items and results	Includes 682 items, such as blood, etc.
	Diseases	Patient's disease, such as cerebral infarction, etc.
Unstructured text data	Patient's readme illness	Patient's readme illness and medical history
	Doctor's records	Doctor's interrogation records

For dataset, according to the different characteristics of the patient and the discussion with doctors, we will focus on the following three datasets to reach a conclusion.

• Structured data (S-data): use the patient's structured data to predict whether the patient is at high-risk of cerebral infarction.

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

• Structured and text data (S&T-data): use the S-data and T-data above to multi-dimensionally fuse the structured data and unstructured text data to predict whether the patient is at high-risk of cerebral infarction.

We introduce the data imputation, CNN-based unimodal disease risk prediction (CNN-MDRP) algorithm. A CNN MDRP method dependent on CNN-UDRP has been proposed .The handling of content information is comparative with CNN-UDRP which can remove 100 highlights about content informational index. For structure information, we extricate 79 highlights. At that point, we direct the component level combination by utilizing 79 highlights in the S information and 100 highlights in T-information. For computation methods, full connection layer are similar with CNNUDRP algorithm. Figure 1 shows the disease prediction model using various classification algorithms. In CNN-MDRP algorithm, there are two divisions of the training process which is elaborated below,

Training parameters of CNN-MDRP: In CNN-MDRP algorithm, the specific training parameters are W1 , W3 new, b 1 , b3 new. They used stochastic gradient method to train parameters, and finally reach the risk assessment of whether the patient suffers from cerebral infarction. Some advanced features shall be tested in future study, such as fractal dimension , biorthogonal wavelet transform etc.

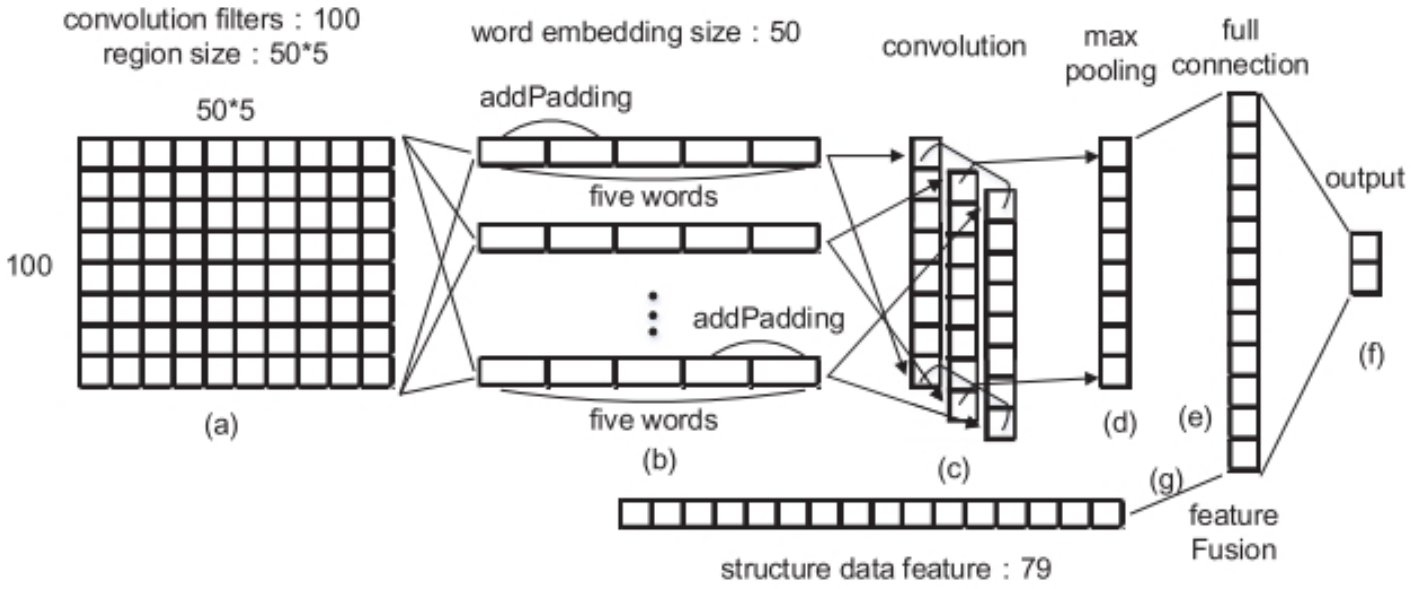

Fig-4.1 CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm.

Training parameters of CNN-MDRP: Stochastic inclination strategy is utilized to prepare parameters to arrive at the hazard appraisal whether the patient experiences cerebral localized necrosis. Some propelled highlights will be tried in future investigation, for example, fractal measurement, orthogonal wavelet change and so forth. The algorithm act as positively with categorical data but poorly if numerical data in the training set.

a) Hospital data: A large volume of datasets of a patient can be given by a hospital which can be processed in the information centre to preserve the patient privacy and authentication of stored data, a security access technique has been created.

b) Structured data: The structured data is nothing but the laboratory data, patient's basic information like patients age, gender, life habits, height, weight etc.

c) Unstructured Data: Unstructured Data is a data of patient's medical history, patient's illness, and doctor's interrogation and diagnosis. The 20 hospitals datasets consisting 20,000 documents and data of patients. The 20 hospital dataset is a popular dataset for experiments in application of a machine learning techniques

For the performance evaluation in the experiment. First, we denote

TP - TRUE POSITIVE - Number of instances correctly predicted as required.

FP - FALSE POSITIVE - Number of instances incorrectly predicted as required.

TN - TRUE NEGATIVE - Number of instances correctly predicted as not required.

FN - FALSE NEGATIVE - Number of instances incorrectly predicted as not required.

Then, we can obtain four measurements: accuracy, precision, recall and F1-measure as follows:

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

4.2 Requirements

The proposed system requires a following configuration for implementation

1. JDK 1.8
2. Database - Mongo DB
3. Server- Apache Tomcat server

4.3 SYSTEM ARCHITECTURE

PROPOSED WORK

In a proposed system we can first get the large volume of a healthcare big data, then that data is considered as training data. Naive Bayes algorithm is used for the clarification of the data. Then after the clarification the hospital data similar type of data can be stored. Then CNN extract the text characteristics automatically. In that we use a CNN MDRP algorithm that uses both structured unstructured hospital data. Selecting the characteristics automatically form a large number of data. This improves the disease prediction rather than previously selected characteristics. CNN- MDRP algorithm helps to accuracy of the result of a disease prediction over a large volume of data from hospital

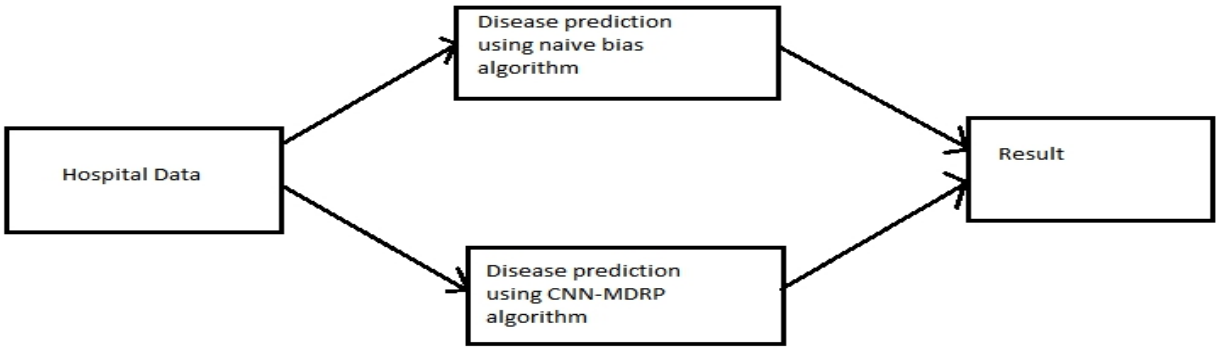

Fig. 4.3 - Block Diagram of Proposed System (Proposed System Architecture)

ALGORITHM Used machine learning algorithm: - Naive Bayes

It is a classification technique based on a Bayes theorem. Javier Bayes algorithm is easy to build and mainly useful for a very large amount of data sets. In a naive Bayes it can convert the data set in a frequency table and then create a likelihood table by finding the probabilities like overcast probability. In our paper we are using the naive Bayes algorithm for the accurate outcome of prediction from the large volume of a medical data.Bayes theorem provides a way of calculating the posterior probability, P(c—x), from P(c), P(x), and P(x—c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence 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 Where C and X are two events (e.g. the probability that the train will arrive on time given that the weather is rainy). Such Nave Bayes classifiers use the probability theory to find the most likely classification of an unseen (unclassified) instance

CHAPTER 5

EXPERIMENTAL RESULTS

According to the discussion before, they got the accuracy,precision, recall, F1-measure and ROC curve under CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms. In this experiment, the selected number of words is 7 and the text feature is 100. As for CNNUDRP (T-data) and CNN-MDRP (S&T data)algorithms, they both run 5 times and seek the average of their evaluation indexes. From the Fig. 8, the accuracy is 0.9420 and the recall is 0.9808 under CNN-UDRP (T-data) algorithm while the accuracy is 0.9480 and the recall is 0.99923 under CNN-MDRP (S&T-data) algorithm. Thus, they draw the conclusion 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.

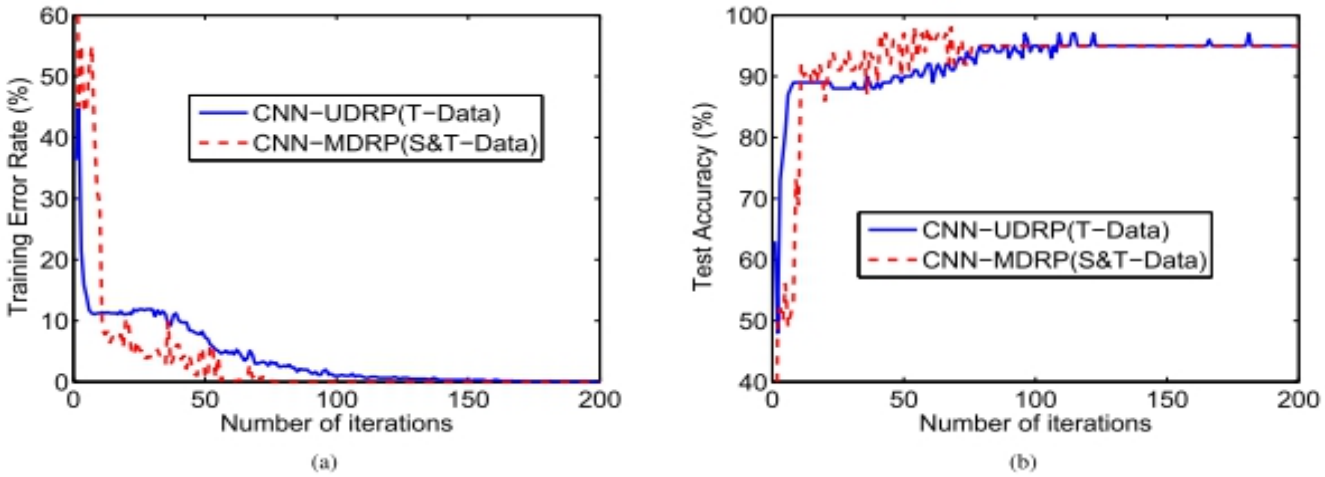

Fig 5 -Effect of iterations on the algorithm. (a) The trend of training error rate with the iterations for CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms. (b) The trend of test accuracy with the iterations for CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms

In summary, the performance of CNN-MDRP (S&T-data) is better than CNNUDRP (T-data). In conclusion, for disease risk modelling, the accuracy of risk prediction depends on the diversity feature of the hospital data, i.e., the better is the feature description of the disease, the higher the accuracy will be. For some simple disease, e.g., hyperlipidemia, only a few features of structured data can get a good description of the disease, resulting in fairly good effect of disease risk prediction. But for a complex disease, such as cerebral infarction mentioned in the paper, only using features of structured data is not a good way to describe the disease. The corresponding accuracy is low, which is roughly around 50%. Therefore, in this paper, they leverage not only the structured data but also the text data of patients based on the proposed CNN-MDPR algorithm. They found that by combining these two

data, the accuracy rate can reach 94.80%, so as to better evaluate the risk of cerebral infarction disease

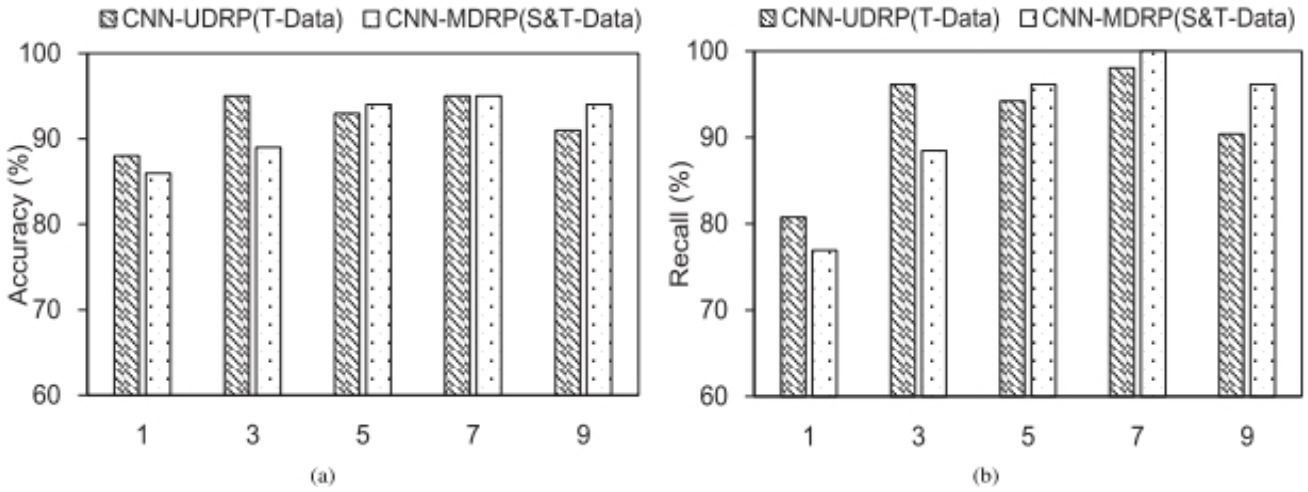

Fig 5.1- Effect of sliding window (word number) in the algorithm. (a) The corresponding accuracy of the CNN-UDRP (T-data) and
CNN-MDRP (S&T-data) algorithms when the number of words for sliding window are 1, 3, 5, 7 and 9. (b) The corresponding recall of the CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms when the number of words for sliding window are 1, 3, 5, 7 and 9.

ROC curve under CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms. In this experiment, the selected number of words is 7 and the text feature is 100. As for CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms, we both run 5 times and seek the average of their evaluation indexes.

From the Fig. 8, the accuracy is 0.9420 and the recall is 0.9808 under CNN-UDRP (T-data) algorithm while the accuracy is 0.9480 and the recall is 0.99923 under CNN-MDRP (S&T-data) algorithm. Thus, we can draw the conclusion 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.

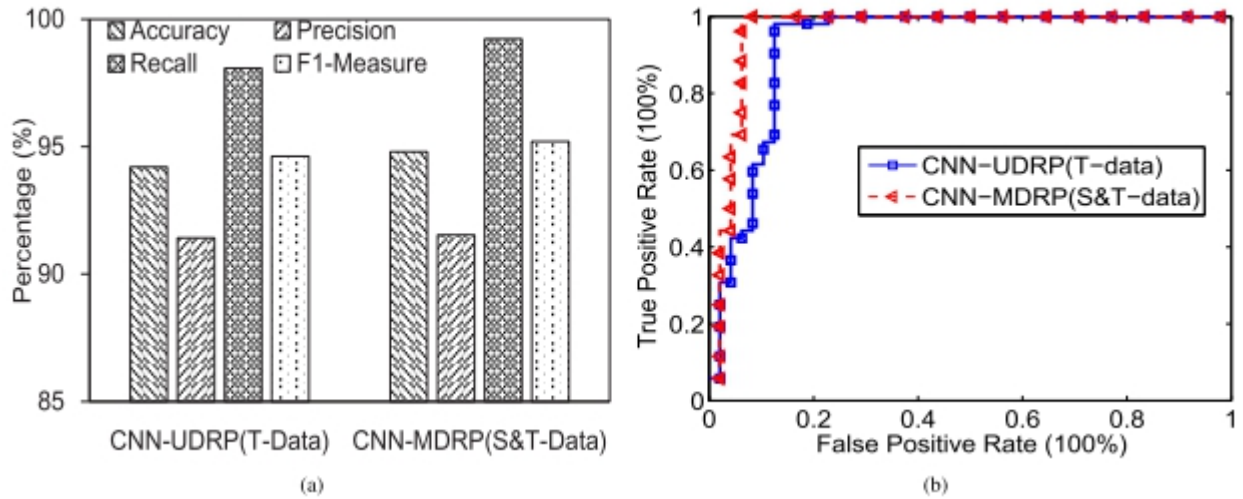

Fig 5.2-Overall results of S&T-data. (a) Comparison of accuracy, precision, recall and F1-measure under CNN-UDRP (T-data) and
CNN-MDRP (S&T-data) algorithms. (b) ROC curves under CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms.

CHAPTER 6

CONCLUSION

In this project, we propose a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm., in CNN-MDRP concentrated on both organized and unstructured information for which the exactness of disease prediction is good and quick when contrasted with the CNN-UDRP. By combining the structured and unstructured data

BIBLIOGRAPHY

1. Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang, Disease Prediction by Machine Learning over Big Data from Healthcare Communities, 2169-3536 (c) 2016 IEEE.

2. Hahab Tayeb*, Matin Pirouz*, Johann Sun1, Kaylee Hall1, Andrew Chang1, Jessica Li1, Connor Song1, Apoorva Chauhan2, MichaelFerra3, Theresa Sager3, Justin Zhan*, Shahram Latifi, Toward Predicting Medical Conditions Using k-Nearest Neighbours, 2017 IEEE International Conference on Big Data.

3. P. Groves, B. Kayyali, D. Knott, and S. van Kuiken, *The'Big Data'Revolution in Healthcare: Accelerating Value and Innovation*. USA: Center for US Health System Reform Business Technology Offifice, 2016.

4. M. Chen, S. Mao, and Y. Liu, ''Big data: A survey,'' *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014.

5. P. B. Jensen, L. J. Jensen, and S. Brunak, ''Mining electronic health records: Towards better research applications and clinical care,'' *Nature Rev. Genet.*, vol. 13, no. 6, pp. 395–405, 2012.

6. D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, ''A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3033–3049, Dec. 2015.

7. M. 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.

8.M. Chen, Y. Ma, J. Song, C. Lai, and B. Hu, ''Smart clothing: Connecting human with clouds and big data for sustainable health monitoring,'' *ACM/Springer Mobile Netw. Appl.*, vol. 21, no. 5, pp. 825–845, 2016.

9. M. Chen, P. Zhou, and G. Fortino, ''Emotion communication system,'' *IEEE Access*, vol. 5, pp. 326–337, 2017, doi: 10.1109/ACCESS.2016.2641480.