A Project Based Seminar Report

on

"Disease Prediction using Machine Learning"

Submitted to the

Savitribai Phule Pune University

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

in

Information Technology

by

Avanti Dorle

T150058534

TE-10

Under the guidance of

Dr. Anant M. Bagade

Department of Information Technology

Pune Institute of Computer Technology

Sr. No 27, Pune-Satara Road, Dhankawadi, Pune - 411 043

2017-2018

CERTIFICATE

This is to certify that the project based seminar report entitled "**Disease Prediction using Machine Learning**" being submitted by **Avanti Dorle (T150058534 / TE-10)** is a record of bonafide work carried out by her under the supervision and guidance of **Dr. Anant M. Bagade** in partial fulfillment of the requirement for **TE (Information Technology Engineering) – 2015 course** of Savitribai Phule Pune University, Pune in the academic year 2017-2018.

Date: 11/04/2018

Place: Pune

Dr. Anant M. Bagade
Guide

Dr. B. A. Sonkamble
Head of the Department

Dr. P. T. Kulkarni
Principal

This Project Based Seminar report has been examined by us as per the Savitribai Phule Pune University, Pune requirements at Pune Institute of Computer Technology, Pune – 411043 on

. . . . . .

Internal Examiner

External Examiner

ACKNOWLEDGEMENT

No work can be considered complete without a word of appreciation to all those who have contributed for it.

We would like to take this opportunity to express our deep sense of humbleness, our sincerity, heart full and profound gratitude towards out Head of Department (Information Technology) Dr. B.A.Sonkamble for providing us opportunity to work on this seminar.

We would also like to thank our guide Dr. Anant M. Bagade for his invaluable advice and wholehearted cooperation without which this project based seminar wouldn't have been a success.

We are thankful to Prof. M. R. Khodaskar and all the faculty members of our department for helping us move forward and providing necessary guidance. We are also thankful to our classmates for their cooperation and moral support.

Last but not least, we thank to all the non-teaching staff of our department, library staff and those who indirectly helped us in completing our seminar.

Avanti Dorle (T150058534)

ABSTRACT

Machine Learning plays a vital role in Medical field for prediction of diseases. We are demonstrating algorithms to predict given four diseases Cancer, Diabetes, Coronary diseases, Cerebral Infarction using machine learning techniques. Cancer can be predicted using Support Vector Machine, K-Nearest Neighbor, Decision Tree, Naive Bayes methods of machine learning of which SVM can be used for better prediction of the above disease. Multi-View Deep Convolutional Neural Network can also be used for predicting Cancer. Diabetes can be predicted using Logistic Regression, Random Forest and Naïve Bayes algorithms of machine learning of which Naive Bayes algorithm can be used for better prediction of the above disease. Coronary diseases can be predicted using J48 algorithm, Naive Bayes algorithm and Random forest algorithm. When comparing the results with Naive Bayes and Random Forest algorithm, J48 algorithm achieved higher precision, recall and F-measure than Naive Bayes and Random Forest algorithm. Cerebral Infarction disease can be predicted using CNN-Based Unimodal Disease Risk Prediction, CNN-Based Multimodal Disease Risk Prediction. Among the algorithms mentioned CNN-Based Multimodal Disease Risk Prediction algorithm provide the best results. Furthermore, we have demonstrated a comparative analysis of various algorithms used for every disease. This analysis gives an insight as to which of the algorithms proves to be more accurate as compared to others.

Keywords: Machine Learning; Convolutional Neural Network; K-Nearest Neighbours; Naive Bayes; Regression; Random Forest

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

The 21st century has been an era of data driven decisions. Machine learning allows building models to quickly analyze data and deliver results, leveraging both historical and real-time data. With machine learning, healthcare service providers can make better decisions on patient's diagnosis and treatment options, which leads to the overall improvement of healthcare services. In this seminar we are presenting Machine Learning algorithms to predict major four diseases (Cancer, Cerebral Infarction, Coronary Heart Disease and Diabetes). Further a comparative analysis of the various algorithms used in prediction of respective diseases is done.

1.2 Motivation

Today large number of deaths are caused due to late diagnosis of diseases such as cancer, coronary diseases, brain tumor, diabetes, and many more. In this project we are trying to predict such diseases with the help of the attributes related to their daily habits, patient's demographics, historical data.

The healthcare industry collects large amounts of health-care data and that need to be mined to discover hidden information for effective decision making. Motivated by the world-wide increasing mortality of various disease patients each year and the availability of huge amount of patients' data from which to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of diseases. According to the research, the algorithms that are demonstrated in this seminar have been proved to be highly accurate and precise. The main motivation behind disease prediction is to improve the accuracy of demonstrated algorithms, by training better models with Machine Learning techniques.

1.3 Aim and Objectives

Aim:

Using Machine Learning techniques and algorithms to train models for predicting 4 major diseases viz. Cancer, Coronary, Diabetes, Cerebral Infarction, with improved accuracy and better precision.

Objectives

- To research on various existing Machine Learning algorithms, and techniques that have been used by researchers for disease prediction.
- To apply the knowledge gained, to train models for the algorithms
- To compare, analyse and verify the accuracy and precision of every algorithm for the respective diseases.
- According to the comparative analysis, selecting the algorithm with most accurate and precise predictions, among all the other algorithms for a particular disease among four of them.
- To try and improve the accuracy of the selected algorithm based on its features, using Machine Learning techniques.

1.4 Introduction to Disease Prediction using ML

Disease prediction has long been regarded as a critical topic. In this project, we are using neural networks and machine learning techniques to solve this problem. We have presented various algorithms for predicting four common diseases (Diabetes, Breast Cancer, Coronary diseases, Cerebral Infarction) and also the comparative analysis of these algorithms are done. Cancer can be predicted using Support Vector Machine, K-Nearest Neighbor, Decision Tree, Naive Bayes methods of machine learning. Coronary diseases can be predicted using J48 algorithm, Naive Bayes algorithm and Random forest algorithm. Diabetes can be predicted using Logistic Regression, Random Forest and Naïve Bayes algorithms. Cerebral Infarction disease can be predicted using CNN-Based techniques.

CHAPTER 2
LITERATURE SURVEY

The artificial neural network (ANN)–a machine learning technique inspired by the human neuronal synapse system–was introduced in the 1950s. However, the ANN was previously limited in its ability to solve actual problems, due to the vanishing gradient and overfitting problems with training of deep architecture, lack of computing power, and primarily the absence of sufficient data to train the computer system. Interest in this concept has lately resurfaced, due to the availability of big data, enhanced computing power with the current graphics processing units, and novel algorithms to train the deep neural network. Recent studies on this technology suggest its potentially to perform better than humans in some visual and auditory recognition tasks, which may portend its applications in medicine and healthcare, especially in medical imaging, in the foreseeable future. According to literature reference [3] article offers perspectives on the history, development, and applications of deep learning technology, particularly regarding its applications in medical imaging.

The modern approach to healthcare is to prevent the disease with early intervention rather than go for a treatment after diagnosis. Traditionally, physicians or doctors use a risk calculator to assess the possibility of disease development. These calculators use fundamental information such as demographics, medical conditions, life routines, and more to calculate the probability of developing a certain disease. Such calculations are done using equation-based mathematical methods and tools. The challenge here is the low accuracy rate with a similar equation-based approach.

For an example, according to study reference [8], can predict the hospitalization with only 56% of accuracy for a long-term cardiovascular disease. Experts in the field are working on the methodologies to identify, develop, and fine-tune machine learning algorithms and models which can deliver accurate predictions.To develop a strong and more accurate machine learning model, we can use data collected from studies carried out, patient demographics, medical health records, and other sources. The difference between traditional and machine learning approach for disease prediction is the number of dependent variables to consider. In a traditional approach, they consider very few variables that you can count on your figure such as age, weight, height, gender and more (due to computational limitation). On the other hand, machine learning being processed on computing devices can consider a large number of variables, which results in a better accuracy of healthcare data.

With big data growth in biomedical and healthcare communities, accurate analysis of medical data benefits early disease detection, patient care, and community services. However, the analysis accuracy is reduced when the quality of medical data is incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. According to the research [1], machine learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities is streamlined. The modified prediction models over real-life hospital data collected from central China in 2013–2015 is experimented. Experiment on a regional chronic disease of cerebral infarction is carried out. A new convolutional neural network (CNN)-based multimodal disease risk prediction algorithm using structured and unstructured data from hospital is proposed. Compared with 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 algorithm. Further studies were being carried out to propose Machine learning and deep learning models that can improve accuracy of the given approach. In previous studies, to overcome the difficulty of incomplete data, a latent factor model is used to reconstruct the missing data. This helps for efficient prediction of diseases.

Artificial intelligence (AI), a computer system aiming to mimic human intelligence, is gaining increasing interest and is being incorporated into many fields, including medicine. Stroke medicine is one such area of application of AI, for improving the accuracy of diagnosis and the quality of patient care. According to literature reference [2] for stroke management, adequate analysis of stroke imaging is crucial. Recently, AI techniques have been applied to decipher the data from stroke imaging and have demonstrated some promising results. In the very near future, such AI techniques may play a pivotal role in determining the therapeutic methods and predicting the prognosis for stroke patients in an individualized manner. In literature [2] it has offer a glimpse at the use of AI in stroke imaging, specifically focusing on its technical principles, clinical application, and future perspectives. Similar techniques can be used for prediction of Cerebral Infarction as well.

CHAPTER 3

METHODOLOGY AND ALGORITHMS

3.1 Data Description

Prediction using traditional disease risk models usually involves a machine learning algorithm (e.g., logistic regression and regression analysis, etc.), and especially a supervised learning algorithm by the use of training data with labels to train the model. In the test set, patients can be classified into groups of either high-risk or low-risk.For dataset, according to the different characteristics of the patient 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.

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

Table 3.1: Item Taxonomy

[Reference 1]

According to the research [1] the data focus on inpatient department data which included 31919 hospitalized patients with 20320848 records in total. The inpatient department data is mainly composed of structured and unstructured text data. The structured data includes laboratory data and the patient's basic information such as the patient's age, gender and life habits, etc. While the unstructured text data includes the patient's narration of his/her illness, the doctor's interrogation records and diagnosis, etc.

3.2 Disease Risk Prediction

According to [1] the basic method for disease risk prediction is considered as follows. The risk prediction model for cerebral infarction is regarded as the supervised learning methods of machine learning, i.e., the input value is the attribute value of the patient, X = (x1, x2, . . . , xn) which includes the patient's personal information such as age, gender, the prevalence of symptoms, and living habits (smoking or not) and other structured data and unstructured data. The output value is C, which indicates whether the patient is amongst the cerebral infarction high-risk population. C = { C0,C1} , where, C0 indicates the patient is at high-risk of cerebral infarction, C1 indicates the patient is at low-risk of cerebral infarction. Machine learning and deep learning algorithms are introduced.

- **For S-data**, three conventional machine learning algorithms are used:
    - Naive Bayesian (NB)
    - K-nearest Neighbour (KNN)
    - Decision Tree (DT)
- **For T-data,**
    - Convolutional Neural Network-based Unimodal Disease Risk Prediction (CNN-UDRP)
- **For S& T data,**
    - Convolutional Neural Network-based Multimodal Disease Risk Prediction (CNN-MDRP)

3.3 Evaluation Methods

- **Performance measures**:
    - **True Positive (TP)**: The number of instances correctly predicted as required
    - **False Positive (FP)**: The number of instances incorrectly predicted as required
    - **True Negative (TN)**: The number of instances correctly predicted as not required
    - **False Negative (FN)**: The number of instances incorrectly predicted as not required
- **Evaluation Methods:**
    - **Accuracy**: A ratio of correctly predicted observations to total observations
    - **Precision**: The ratio of correctly predicted positive observations to the total predicted positive observations.
    - **Recall**: The ratio of correctly predicted positive observations to all observations in actual class.
    - **F1 score**: The weighted average of precision and recall

3.4 Methods

3.3.1 CNN-based Unimodal Disease Risk Prediction

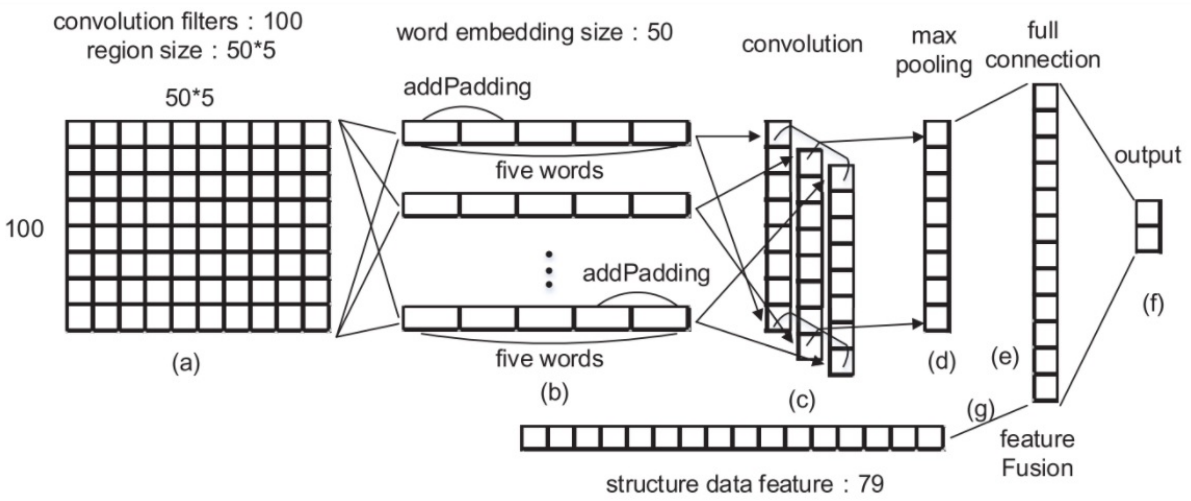

Figure 3.1: CNN based MDRP

According to reference, for the processing of medical text data, CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm is utilized which can be divided into the following five steps.

1) **Representation of Text data**

   As for each word in the medical text,the distributed representation of Word Embedding in natural language processing is used, i.e. the text is represented in the form of vector.

2) **Convolution Layer of Text CNN**

   Two words from the front and back is chosen of each word vector in the text, i.e. use the row vector as the representation, to consist a $50 \times 5 = 250$ row vector. The selected weight matrix is as shown in Fig. 3.1(a), i.e., weight matrix includes 100 convolution filters and the size of each filter regions is 250.

3) **Pool Layer of Text CNN**

Taking the output of convolution layer as the input of pooling layer, we use the max pooling (1-max pooling) operation as shown in Fig. 3.1(d), i.e., select the max value of the n elements of each row in feature graph matrix. In spite of different length of the input training set samples,the text is converted into a fixed length vector after convolution layer and pooling layer.

4) **Full Connection Layer of Text CNN**

Pooling layer is connected with a fully connected neural network as shown in Fig. 3.1(e).

5) **CNN Classifier**

The full connection layer links to a classifier, for the classifier, we choose a softmax classifier, as shown in Fig. 3.1(f).

3.3.2 CNN-based Multimodal Disease Risk Prediction

From what has been discussed above, we can get the information that CNN-UDRP only uses the text data to predict whether the patient is at high risk of cerebral infarction. As for structured and unstructured text data, we design a CNN-MDRP algorithm based on CNN-UDRP as shown in Fig. 3.1. The processing of text data is similar with CNN-UDRP, as shown in Fig. 3.1(a-d), which can extract 100 features about text data set. For structure data, we extract 79 features. Then, we conduct the feature level fusion by using 79 features in the S-data and 100 features in T-data, as shown in Fig. 3.1(g). For full connection layer, computation methods are similar with CNN-UDRP algorithm. Since the variation of features number, the corresponding weight matrix and bias change to new weight and bias. The training is as follows:

1) **Training Word Embedding**

Word vector training requires pure corpus, the purer the better, that is, it is better to use a professional corpus. In reference with [1], text data of all patients in the hospital from the medical large data center is extracted. Using word2vec tool n-skip gram algorithm trains the word vector, word vector dimension is set to 50, after training we get about 52100 words in the word vector.

2) **Training Parameters of CNN-MDRP**

In CNN-MDRP algorithm, the specific training parameters stochastic gradient method is used to train parameters, and finally reach the risk assessment of whether the patient suffers from cerebral infarction.

3.4 Results

Results are evaluated based on following features:

Running time:

We can see the running time of CNN-UDRP (T-data) and CNN-MDRP (S& T-data) are basically the same from the Fig. 3.2, i.e. although the number of CNN-MDRP (S& T-data) features increase after adding structured data, it does not make a significant change in time.

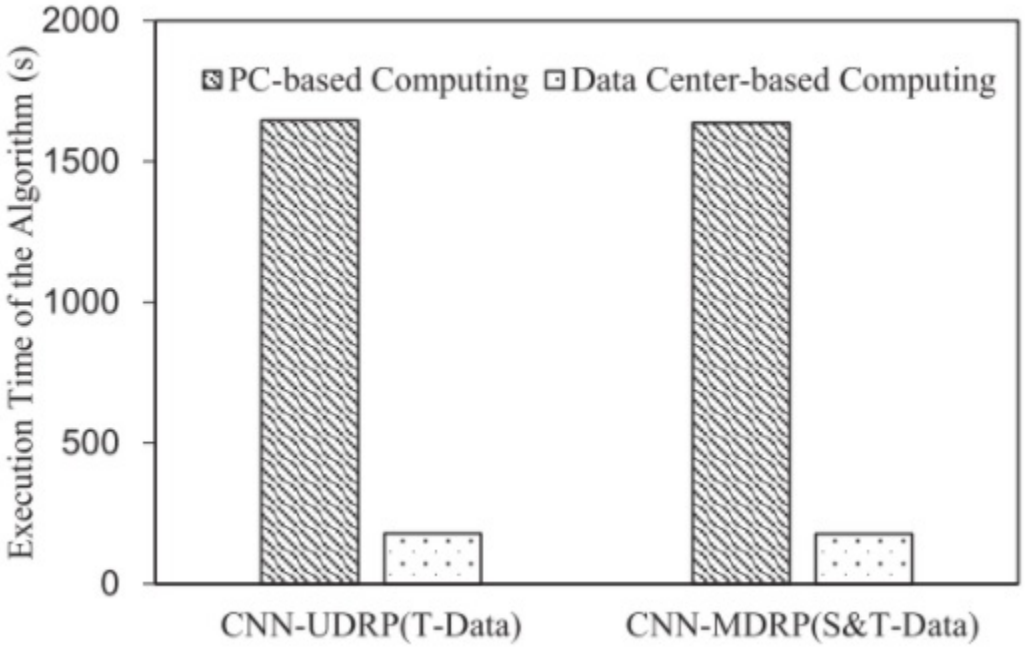

Figure 3.2: Running Time Comparison

[Reference 1]

Effect of Iterations

In Fig. 3.3, we can obtain when the number of iterations are 70, the training process of CNN-MDRP (S& T-data) algorithm is already stable while the CNN-UDRP (T-data) algorithm is still not stable. In other words, the training time of MDRP (S& Tdata) algorithm is shorter, i.e. the convergence speed of CNN-MDRP (S& T-data) algorithm is faster.

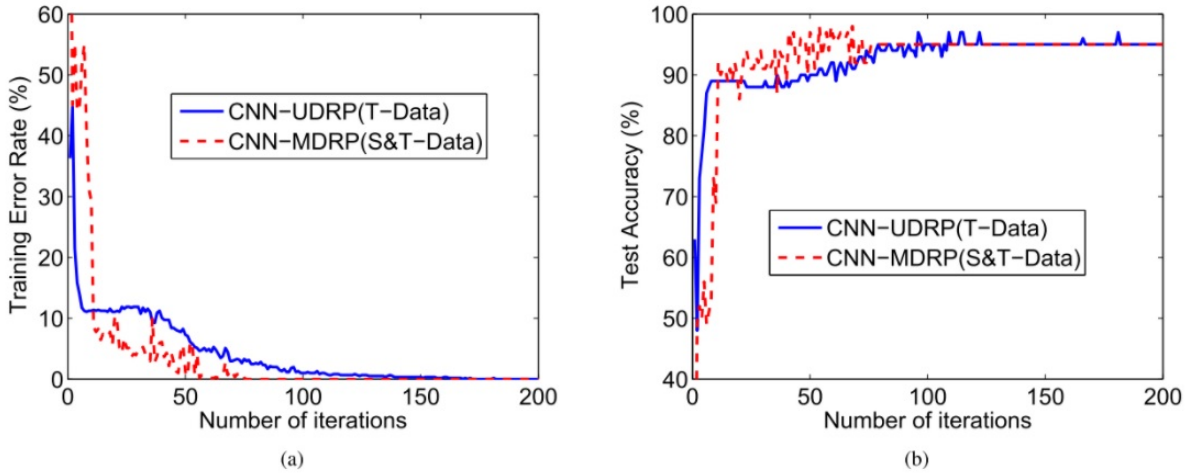

Figure 3.3: Number of Iterations Comparison

[Reference 1]

Effect of Text features

Fig. 3.4 shows the accuracy and recall of each feature after it go through 200 times of iteration. From the Fig. 3.4(a) and Fig. 3.4(b), when the feature number of text is smaller than 30, the accuracy and recall of CNN-UDRP (T-data) and CNN-MDRP (S& T-data) algorithms are smaller than the feature number of text is bigger than 30 obviously. the accuracy of CNN-MDRP (S& T-data) algorithm is more stable than CNN-UDRP (T-data) algorithm, i.e. the CNN-MDRP (S& T-data) algorithm is reduced fluctuation after adding structured data. As shown in Fig. 3.4(b), after adding structured data, the recall of CNN-MDRP (S& T-data) algorithm is higher than CNN-UDRP (T-data) algorithm obviously. This shows that the recall of algorithm is improved after adding structured data.

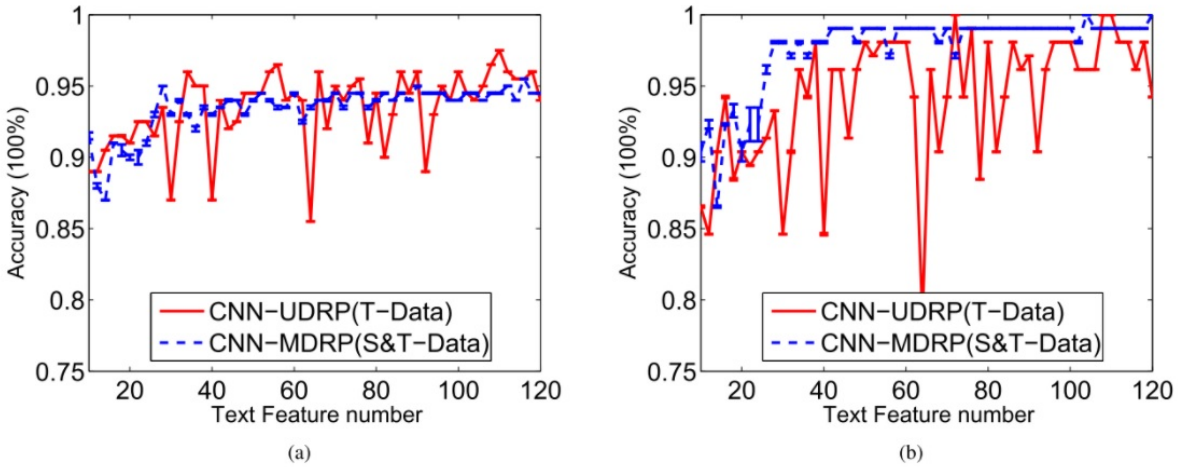

Figure 3.4: Comparison based on text features

[Reference 1]

Overall results:

According to the reference [1], 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, both run 5 times and seek the average of their evaluation indexes. From the Fig. 5, 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, conclusion can be drawn 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).

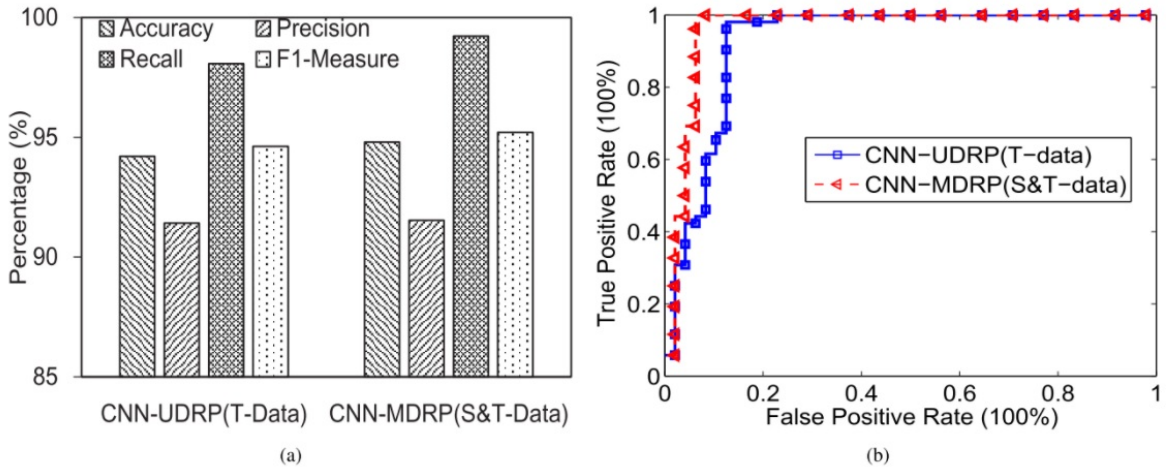

Figure 3.5: Overall comparison

[Reference 1]

CHAPTER 4

ADVANTAGES AND DISADVANTAGES

4.1 Advantages

- The approach presented above gives an insight into how three types of data can be used for accurate prediction of Cerebral Infarction.
- This approach not only focuses on structured data but also textual and pictorial data.
- CNN usually uses images for prediction or recognition, thus, the results will be more accurate in case of data in the form of images.

4.2 Disadvantages

- When considering image data, this approach might not give 100% accurate results which is risky as far as Cerebral Infarction is concerned

CHAPTER 5

APPLICATIONS

- This study can be used for further enhancements in Medical fields
- The predictions done prior to the diagnosis of disease is beneficial for the Medical sector as well as for the patients
- This method can be used in remote and rural areas lagging with hospitals and medications. Prior prediction can help rural population to take preventive measures.

CHAPTER 6

ENHANCEMENTS

The method and approach described above, gives highly accurate results for disease predictions. Although they are not completely accurate. There still exists some probability that the results are incorrect. This can cause serious damage to the patient as well as the Medical sector. Future enhancements are to train various models, based on different variations of features, types of data viz. structured, unstructured, textual data, image data etc, and more demographics of patients, as well as historical data. To train models for even better results and improved accuracy.

CHAPTER 7

CONCLUSION

According to the study, we have presented a seminar on Disease Prediction using Machine Learning. In this seminar we demonstrated four major diseases Cerebral Infarction, Coronary disease, Diabetes, and Cancer. As these diseases are critical, we have studied the ways to predict them before they are diagnosed. Using various machine learning algorithms and techniques, we conclude that there are algorithms that prove to be the most accurate among other demonstrated algorithms. Cerebral Infarction can be predicted using CNN based algorithms like, CNN based Unimodal Disease Risk Prediction, and CNN based Multimodal Disease Risk Prediction, among which CNN-MDRP proves to be the most accurate with 94% accuracy and convergence speed faster than CNN-UDRP.

CHAPTER 8

REFERENCES

[1] Min Chen; Yixhue Hao; Kai Hwang; Lu Wang; Lin Wang "Disease Prediction by Machine Learning Over Big Data from Healthcare Communities" (Vol. 5 IEEE Access ISSN: 2169-3536)

[2] Eun-Jae Lee,a∗ Yong-Hwan Kim,a∗ Namkug Kim,b Dong-Wha Kanga "Deep into the Brain: Artificial Intelligence in Stroke Imaging" (Journal of Stroke 2017;19(3):277-285)

[3] June-Goo Lee, PhD,1 Sanghoon Jun, PhD,2,3 Young-Won Cho, MS,2,3 Hyunna Lee, PhD,2,3 Guk Bae Kim, PhD,2,3 Joon Beom Seo, MD, PhD,2,∗ and Namkug Kim, PhD2,3,∗ "Deep Learning in Medical Imaging: General Overview" (Korean J Radiol. 2017 Jul-Aug 18(4): 570-584 doi: 10.3348/kjr.2017.18.4.570)

[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] 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 Office, 2016.

[8] https://hbr.org/2017/05/how-machine-learning-is-helping-us-predict-heart-disease-and-diabetes