Disease Prediction Using Patient Treatment History and Health Data

Problem Statement

Healthcare industry has become big business. The healthcare industry produces large amounts of health-care data daily that can be used to extract information for predicting disease that can happen to a patient in future while using the treatment history and health data.This hidden information in the healthcare data will be later used for affective decision making for patient's health. Also, this area need improvement by using the informative data in healthcare.

Major challenge is how to extract the information from these data because the amount is very large so some data mining and machine learning techniques can be used. Also,the expected outcome and scope of this project is that if disease can be predicted than early treatment can be given to the patients which can reduce the risk of life and save life of patientsand cost to get treatment of diseases can be reduced up to some extent by early recognition. For this problem, aprobabilistic modeling and deep learning approach will train a Long Short-Term Memory (LSTM) recurrent neural network and two convolutional neural networks for prediction of disease.

The rapid adoption of electronic health records has created a wealth of new data about patients, which is a goldmine for improving the understanding of human health. The above method isused to predict diseases using patient treatment history and health data.

Background

Disease prediction using patient treatment history and health data by applying data mining and machine learning techniques isongoing struggle for the past decades. Many works have been applied data mining techniques to pathological data or medical profiles for prediction of specific diseases. These approaches tried to predict the reoccurrence of disease. Also, some approaches try to do prediction on control and progression of disease. The recent success of deep learning in disparate areas of machine learning has driven a shift towards machine learning models that can learn rich, hierarchical representations of raw data with little preprocessing and produce more accurate results. Numbers of papers have been published on several data mining techniques for diagnosis of heart disease such as Decision Tree, Naive Bayes, neural network, kernel density, automatically defined groups, bagging algorithm and support vector machine showing different levels of accuracies in diseases prediction.In this type of research generally used tool is Waikato Environment for Knowledge Analysis (WEKA).

Methodology

Step 1: Data collection and dataset preparation

This will involve collection of medical information artifacts from various sources like hospitals, discharge slips of patientsand from UCI repository then preprocessing is applied on dataset which will remove all the unnecessary data and extract important features from data.

Step 2: Developing a probabilistic modeling and deep learning approach (RNN)for Disease Prediction

In this step probabilistic modeling and deep learning approach based on RNN is to be developed it will runs effectively on extensive databases of healthcare. And generate decision tree also it can deal with a huge number of information variables without variable deletion.

Step 3: Training and experimentation on datasets

The Disease Predictionmodel will be trained on the dataset of diseases to do the prediction accurately and produce Confusion matrix.

Step 4: Deployment and analysis on real life scenario

The trained and tested prediction model will be deployed in a real-life scenario made by the human experts & will be leveraged for further improvement in the methodology and will follow the above architecture.

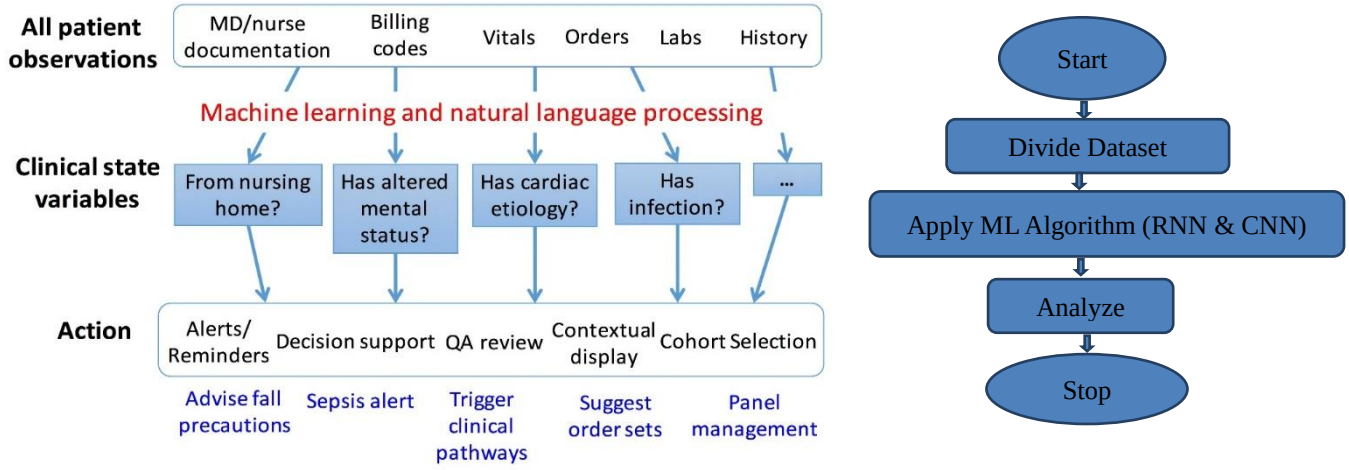

Figure.1 Workflow of diseases prediction system using health data and history of treatment[1]

Experimental Design

*Dataset:*Heart Disease Data Set is available at UCI which is Machine Learning Repository[2], Prime Indians Diabetes Dataset is available on KAGGLE[3], Breast Cancer dataset is available at UCI[4] and many other health related dataset are available on UCI,Heidelberg University Hospital has 27,000 fully anonymized, real-world discharge letters dataset provided by them on request which can be used for experimentation and evaluation**.**

*Evaluation Measures:*For measuring the accuracy or effectiveness of the implemented system various metricshave been proposed such as Absolute Error rate (AER),Accuracy v/s number of observation in terms of diseases prediction model will be measured. This will help us in prediction of diseases.

Software and Hardware Requirements

Python based Deep Learning libraries will be exploited for the development and experimentation of the project. Tools such as Anaconda Python, and python libraries will be utilized for this process. Training will be conducted on NVIDIA GPUs for training aprobabilistic modeling and deep learning approach for diseases prediction. We can use medical hardware devices for capturing the real data or test the results on real-time data.

[1]http://clinicalml.org/research.html
[2]http://archive.ics.uci.edu/ml/datasets/Heart+Disease
[3]https://www.kaggle.com/uciml/pima-indians-diabetes-database/data
[4]https://archive.ics.uci.edu/ml/datasets/Breast+Cancer