AI-Assisted Prediction on Potential Health Risks With Regular Physical Examination Records

Rishee Raj[1], Rupam Paul[2], SachinDolta[3], Shivam Pandey[4], Gayathri R.[5]
[1, 2, 3, 4] Dept of Information Science & Engineering
[5]Assistant Professor, Dept of Information Science & Engineering
[1, 2, 3, 4, 5] Sapthagiri College of Engineering, Chikkasandra, Bangalore, Karnataka, India

Abstract- *The amount data in the health industry is increasing rapidly and is expected to increase drastically in coming years. Healthcare services although equipped with modern technologies for remedial the diseases grapple when it comes to preventing the diseases beforehand. Adoption of Machine Learning solutions will play an important role in transforming the outcomes of the healthcare industry by promoting facts based analysis and providing patient-centric manner. In this age of Technology, we can provide solutions to identify individuals who are prone to certain lifestyle diseases. Think of identifying an individual having an increased risk of diabetes after 10 years, now. With the advent of new data analysis equipment and technologies, such analytical systems can be designed which can identify individuals with increased risk. This document provides an overview of data analytics, different technologies that can be used in data and its force on this field to make some useful predictions based upon analyzing a variety of datasets. Finally, we provide a model which can be used for predictive analytics using machine learning algorithms to predict the chances of a person to be prone to a disease.*

Key words: *Machine Learning, Healthcare, Prediction System*

I. INTRODUCTION

Predicting the risk of potential diseases from Patient Health Records (PHR) has attracted considerable attention in recent years, especially with the development of machine learning techniques. Compared with traditional machine learning models, deep learning-based approaches achieve superior performance on risk prediction task. However, none of existing work explicitly takes prior medical knowledge (such as the relationships between diseases and corresponding risk factors) into account. In medical domain, knowledge is usually represented by discrete and arbitrary rules. Thus, how to integrate such medical rules into existing risk prediction models to improve the performance is a challenge. To tackle this challenge, we propose a novel and general framework for risk prediction task, which can successfully incorporate discrete prior medical knowledge into all of the state-of-the-art predictive models using posterior regularization technique. Different from traditional posterior regularization, we do not need to manually set a bound for each piece of prior medical knowledge when modeling desired distribution of the target disease on patients. Moreover, the proposed system can automatically learn the importance of different prior knowledge with a log-linear model. Experimental results on three real medical datasets demonstrate the effectiveness of the proposed framework for the task of risk prediction.

There are several advantages to quantitative prediction tools that accurately foretell the occurrence of a disease, its prognosis or course, or an individual's likelihood to respond to a certain treatment. Such tools

a) enable patients and their families to make more informed decisions about treatment and prevention (for instance, balancing the side-effects of a prevention regimen against the individual's likelihood of experiencing that outcome);
b) help clinicians precisely tailor care by planning treatment and prevention; and
c) aid health care systems in allocating resources to patients most at risk for an outcome. The latter process is known in medicine as risk stratification, or the ordering large numbers of patients in strata reflecting increasing levels of risk for the health outcome.

1.2 Aim & Objective: -

1. The Main concept is to determine medical diseases according to given symptoms and daily routine and when user search the hospital, the hospital which is nearest to their current location is given.
2. Determine medical diseases according to given symptoms & daily Routine.
3. Prediction is done on the word count, laboratory results and diagnostic data.

II. LITERATURE REVIEW

1]. "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks" Author,- Srinivas K, Rani B

K, Govardhan A. The healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor'. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. In this study, we briefly examine the potential use of classification based data mining techniques such as rule based, decision tree, naïve bayes and artificial neural network to massive volume of healthcare data. For data preprocessing and effective decision making One Dependency Augmented Naïve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2) are used. This is an extension of naïve Bayes to imprecise probabilities that aims at delivering robust classifications also when dealing with small or incomplete data sets. Discovery of hidden patterns and relationships often goes unexploited. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established.

Disadvantage: -

- For predicting heart attack significantly 15 attributes are listed
- Besides the 15 listed in medical literature we can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules.
- Categorical data is used •Text mining is not used for of unstructured data.

2]. "Grand challenges in clinical decision support" Author-Sittig D, Wright A, Osheroff J, et al. There is a pressing need for high-quality, effective means of designing, developing, presenting, implementing, evaluating, and maintaining all types of clinical decision support capabilities for clinicians, patients and consumers. Using an iterative, consensusbuilding process we identified a rank-ordered list of the top 10 grand challenges in clinical decision support. This list was created to educate and inspire researchers, developers, funders, and policy-makers. The list of challenges in order of importance that they be solved if patients and organizations are to begin realizing the fullest benefits possible of these systems consists of: improve the human–computer interface; disseminate best practices in CDS design, development, and implementation; summarize patient-level information; prioritize and filter recommendations to the user; create an architecture for sharing executable CDS modules and services; combine recommendations for patients with comorbidities; prioritize CDS content development and implementation; create internet-accessible clinical decision support repositories; use free text information to drive clinical decision support; mine large clinical databases to create new CDS. Identification of solutions to these challenges is critical if clinical decision support is to achieve its potential and improve the quality, safety and efficiency of healthcare. Disadvantage: -

- Identification of solutions to these challenges is critical if clinical decision support is to achieve its potential and improve the quality, safety and efficiency of healthcare.

3]. "Using Electronic Health Records for Surgical Quality Improvement in the Era of Big Data" Author Anderson J E, Chang D C. Many healthcare facilities enforce security on their electronic health records (EHRs) through a corrective mechanism: some staff nominally have almost unrestricted access to the records,but there is a strict ex post facto audit process for inappropriate accesses, i.e., accesses that violate the facility's security and privacy policies. This process is inefficient, as each suspicious access has to be reviewed by a security expert, and is purely retrospective, as it occurs after damage may have been incurred.

This motivates automated approaches based on machine learning using historical data. Previous attempts at such a system have successfully applied supervised learning models to this end, such as SVMs and logistic regression. While providing benefits over manual auditing, these approaches ignore the identity of the users and patients involved in a record access. Therefore, they cannot exploit the fact that a patient whose record was previously involved in a violation has an increased risk of being involved in a future violation. Motivated by this, in this paper, we propose a collaborative filtering inspired approach to predicting inappropriate accesses. Our solution integrates both explicit and latent features for staff and patients, the latter acting as a personalized "finger-print" based on historical access patterns. The proposed method, when applied to real EHR access data from two tertiary hospitals and a file-access dataset from Amazon, shows not only significantly improved performance compared to existing methods, but also provides insights as to what indicates an inappropriate access.

4]. "In the paper "Smart health prediction system using data mining"[4] the author has discussed many topics related to data mining techniques such as Naive Bayes, KDD(Knowledge discovery in Database). The Bayesian statistics can be applied to economic sociology and other fields. This checks the patients at initial level and automatically suggest the possible diseases. The system uses

Naive Bayes classifier for the construction of the prediction system. The advantage of this system is that the initial consultation cost of doctor fees can be avoided. Eclipse IDE is used for creating the front end Graphical User Interface and Navicat MySQL is used for backend database purpose. Here java is used as a programming language to connect the database and GUI purpose. The only disadvantage of the system the efficiency in detecting the symptoms or symptom mapping.

5]. "A Smart Health Prediction Using Data Mining" [5] is explaining the similar topics to the paper [4]. But there is detailed explanation of the internal algorithms used in the system. The Naive Bayes algorithm can be used for developing models that are used to assign class labels of different format. Naive Bayes algorithm is not a single, but a group of algorithm based on common principle. The steps involved in the Naive Bayes algorithm include (i) Division of segments, (ii) Comparing the first character of pattern until match occurs, (iii) Comparing the last character of pattern, (iv) Perform each character comparison. Also the hardware requirements used are processor of 2.0 GHZ and Ram of 2GB. The software requirements are JAVA programming language, Mysql 5.0 database and Tomcat server

III. PROPOSED SYSTEM

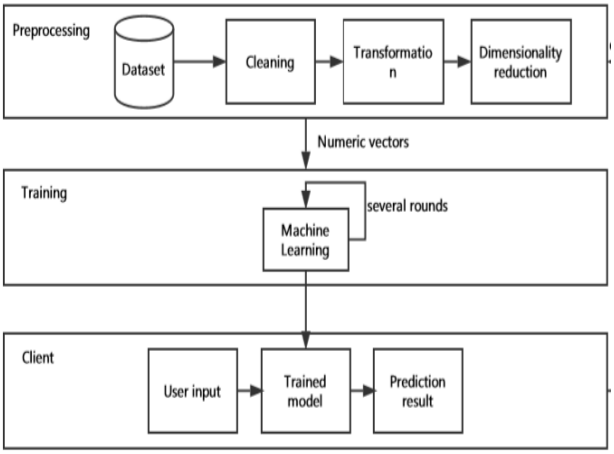

Fig.1:System-Architecture

In this paper, we have combined the structure and unstructured data in healthcare fields that let us assess the risk of disease. The approach of the latent factor model for reconstructing the missing data in medical records which are collected from the hospital. And by using statistical knowledge, we could determine the major chronic diseases in a particular region and in particular community. To handle structured data, we consult hospital experts to know useful features. In the case of unstructured text data, we select the features automatically with the help of k-NN algorithm. We propose a k-NN algorithm for both structured and unstructured data.

The Main Concept to determine medical diseases according to given symptoms & daily Routine when User search the hospital then given the nearest hospital of their current location. The system provides a user-friendly interface for examinees and doctors. Examinees can know their symptoms which accrued in body which set as the while doctors can get a set of examinees with potential risk. A feedback mechanism could save manpower and improve performance of system automatically. The doctor could fix prediction result through an interface, which will collect doctors' input as new training data. An extra training process will be triggered everyday using these data. Thus, our system could improve the performance of prediction model automatically.

The system aims at bridging the gap between medical patients and the availability of doctors and hospitals by using this system to diagnose diseases from the user's symptoms. The project used machine learning algorithms to come up with the probability of diseases. To do so, it takes into account the user's symptoms. This data would then be analyzed and mined through data sets through the use of algorithms such as Naïve Bayes. Moreover, the system would also allow doctors to monitor their patients through the system, without needing to be in physical or even geographical proximity. The project uses Java,MySQL, HTML, CSS, JavaScript, and Jquery

Advantages are:

- Increases human-computer interactions
- Location of User is detected.
- Recommended the hospital and doctor to patient according to diseases Predicted.
- Provided medicine for diseases which is predicted .
- Fast Prediction system.
- Scalable, Low-cost.
- Comparable quality to experts.

IV. METHODOLOGY

The main objective of this project is to develop an application which uses some ML algorithms, languages used will be java. The system takes the symptoms from the users which they are feeling at that moment and runs a Machine Learning algorithm in the DB to detect the disease from which the user may be suffering. The System collects raw data from the user or consumer. As the massive amount of information is already available from healthcare websites, patients can easily

compare the diagnosis done by their doctors and the related information which is already present on the internet. Also by accessing online support group chat system patients can communicate with other patients who are suffering from similar kinds of diseases, this way they can exchange information who might have suffered the same kind of symptoms. The system uses the provided data from the user and matches the symptoms already stored in the database. The database uses various data mining techniques and an intelligent algorithm.

V. MODULE DESCRIPTION

A. Data Gathering

The raw data is gathered from websites like mayoclinic.org, dataworld.org, and kaggle.com. The raw dataset which is a CSV file contains two columns the first one is disease and the other one is the related symptoms for that particular disease. Every disease contains at least 4-5 symptoms for the same and the data is then sent for the pre-processing so that the python code can be implemented .

The disease symptom dataset contains 2525 rows contains multiple diseases that counts upto 52 unique disease and for the drug dataset each disease has multiple drug which differ in rating and user count.

Disease	Symptoms
Influenza	Fever, Chill, Headache, Sneeze
Asthma	Wheezing, Cough, Pleuritic pain
Gastritis	Vomiting, Abdominal pain, Intoxication
HIV	Fever, Cough, Diarrhea

Table [1]: Data Collected

B. Data Preprocessing

After gathering the data in raw form the data is transformed into another .csv file which has indexing of every symptom mapped with every disease in 0 and 1's by the method of one-hot encoding. The diseases are made as rows and all the symptoms are made as columns like the table made below, the presence of every symptom for a particular disease is marked as 1 and its absence is marked as 0 all according to the dataset which we collected earlier i.e raw dataset. From table [2] we can infer that the symptoms like fever, chill are present in Influenza that's why it is marked 1 and since Cough and Nausea are not present so it is marked as 0 according to the dataset which we collected same goes with

Disease	Fever	Chill	Nausea	Vomiting
Influenza	1	1	0	0
Asthma	0	0	1	0
Diabetes	0	0	0	1

Asthma and Diabetes as well.

Table [2]: Cleaned Data

C. Algorithm Implementation

Data mining combines fact-based examination, machine learning, and database design to remove shrouded connections from vast databases.It uses two methodologies: Supervised learning and Unsupervised learning. In supervised, a training set is used to display parameters and in unsupervised learning no training set is utilized. In data mining each technique serves a different purpose depending on the objective of the modeling. The two most common modeling objectives are Classification and Prediction. We used a bunch of different algorithms for the diagnosis of the disease like the, Naïve Bayes and KNN. For these classification models we created all combinations of disease and symptom as we have one to one mapping of the disease and symptom, this process is not requisite in Apriori as it does that during its implementation. Due to this, we conclude that the association rules like APRIORI, CDA, etc. work more efficiently

We select several different algorithms and compare their performance in our experiments.

Naive Bayes : This algorithm forms the basis of different data mining and machine learning models. Naïve Bayes works on prediction modules to help predict the possible outcomes. We used python language for the implementation of it with the dataset we collected. Since it is present in package sklearn kit so importing and loading that is mandatory. Then the dataset is been read using read.csv() then 75% of the data is been trained and the remaining 25% is tested against it all of this done is using NaiveBayes() and predict() present in sklearn package and the predicted output is noted but the predicted data sometimes is correct and sometimes isn't as it doesn't know about the values which didn't occur in that 75% of the data, the accuracy is more when the whole dataset is trained. Accuracy Achieved: 80-85%.

KNN Classifier:

K nearest neighbors is a prediction algorithm that works on the Euclidean distance concept. The distance is measured from different clusters the closer the distance the most likely the element is to belong to that cluster. We implemented this in python the dataset is been read using read.csv() and then the class library is loaded since KNN is present in it using library(class) and then the use of functions like model.train() which is used to train the dataset and the prediction model is made using KNeighborsClassifier() and the accuracy is seen, thus the prediction is checked. Accuracy achieved: 75-85%.

V. CONCLUSION

Our system is a self-service system, which can provide personalized healthcare service to examinees with few maintenance personnel. It is a good solution for the mismatch of in sufficient experienced doctors and rising medical demands. It will gather more training data and improve precision automatically, which releases huge amount of manpower and contains great potential for application.

This paper gave a diagram of utilization of information mining procedures in regulatory, clinical, inquire about, furthermore, instructive parts of Clinical Predictions. This paper set up that while the current down to earth utilization of information mining in wellbeing related issues is constrained, there exists an extraordinary potential for information mining systems to enhance different parts of Clinical Predictions. Besides, the inescapable ascent of clinical information will build the potential for information mining systems that enhances the quality and reduces cost of social insurance.

This system has large scope as it has the following features which are:

- Automation of Disease Diagnosis.
- Paper free work helping the environment.
- To increase the efficiency, accuracy for the patients to help them in future.
- Managing the information related to diseases.

VI. ACKNOWLEDGEMENT

I express my sincere gratitude towards my guide of Prof. Gayathri R. for their constant help, encouragement and inspiration throughout the project work. Also I would like to thank the Head of Information Science and Engineering Department Dr. H. R. Ranganathan for him valuable guidance , ability to motive me and even willingness to solve difficulty made it possible to make my project unique and made task easier. My sincere thanks to Principal, Dr. H. Ramakrishna for providing me necessary facility to carry out the work.

REFERENCES

[1] Srinivas K, Rani B K, Govrdhan A. Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks[J]. International Journal on Computer Science & Engineering, 2010, 2(2):250-255.

[2] Delen, D., &Demirkan, H. Data, information and analytics as services. Decision Support Systems, 2013: 55, 359363.

[3] Malik M M, Abdallah S, AlaRaj M. Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review[J]. Annals of Operations Research, 2016:1-26.

[4] Sittig D, Wright A, Osheroff J, et al. Grand challenges in clinical decision support.[J]. Journal of Biomedical Informatics, 2008, 41(2):387.

[5] Anderson J E, Chang D C. Using Electronic Health Records for Surgical Quality Improvement in the Era of Big Data[J]. Jama Surgery, 2015, 150(1):1-6.

[6] Gheorghe M, Petre R. Integrating Data Mining Techniques into Telemedicine Systems[J]. Informatica Economica Journal, 2014, 18(1):120-130.

[7] Kontio E, Airola A, Pahikkala T, et al. Predicting patient acuity from electronic patient records[J]. Journal of Biomedical Informatics, 2014, 51:35-40.

[8] Amarasingham R, Patzer R E, Huesch M, et al. Implementing electronic health care predictive analytics: considerations and challenges.[J]. Health Aff, 2014, 33(7):1148-1154.

[9] Koh H C, Tan G. Data mining applications in healthcare.[J]. Journal of Healthcare Information Management Jhim, 2005, 19(2):64-72.

[10] Menon A K, Jiang X, Kim J, et al. Detecting Inappropriate Access to Electronic Health Records Using Collaborative Filtering[J]. Machine Learning, 2014, 95(1):87-101.

[11] Yoo I, Alafaireet P, Marinov M, et al. Data mining in healthcare and biomedicine: a survey of the literature.[J]. Journal of Medical Systems, 2012, 36(4):2431-2448.

[12] Miller R A. Medical Diagnostic Decision Support SystemsPast, Present, And Future A Threaded Bibliography and Brief Commentary[J]. Journal of the American Medical Informatics Association Jamia, 1994, 1(1):8.

[13] West D, West V. Model selection for a medical diagnostic decision support system: a breast cancer detection case.[J]. Artificial Intelligence in Medicine, 2000, 20(3):183-204.

[14] Song J H, Venkatesh S S, Conant E A, et al. Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses.[J]. Academic Radiology, 2005, 12(4):487-95.

[15] Nattkemper T W, Arnrich B, Lichte O, et al. Evaluation of radiologicalfeaturesforbreasttumourclassificationinclinical screeningwithmachine learning methods[J]. Artificial Intelligence in Medicine, 2005, 34(2):129139.

[16] Nattkemper T W, Wismller A. Tumor feature visualization with unsupervised learning[J]. Medical Image Analysis, 2005, 9(4):344.

[17] Iakovidis D K, Maroulis D E, Karkanis S A. An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy[J]. Computers in Biology & Medicine, 2006, 36(10):1084-1103.

[18] Billah M, Waheed S, Rahman M M. An Automatic Gastrointestinal Polyp Detection System in Video Endoscopy Using Fusion of Color Wavelet and Convolutional Neural Network Features.[J]. International Journal of Biomedical Imaging,2017,(2017-8-14), 2017, 2017(1):1-9.

[19] Jiang T, Qian S, Hailey D, et al. Text Data Mining of Aged Care Accreditation Reports to Identify Risk Factors in Medication Management in Australian Residential Aged Care Homes[J]. Studies in Health Technology & Informatics, 2017, 245:892.

[20] Seokho Kang, Personalized prediction of drug efficacy for diabetes treatment via patient-level sequential modelin[J]. Artificial Intelligence in Medicine, 2018.

[21] Koh H C, Tan G. Data mining applications in healthcare[J]. J Healthc Inf Manag, 2005, 19(2):64-72.

[22] Bozcuk H, Bilge U, Koyuncu E, et al. An application of a genetic algorithm in conjunction with other data mining methods for estimating outcomeafterhospitalizationincancerpatients.[J].MedicalScienceMonitor International Medical Journal of Experimental & Clinical Research, 2004, 10(6):CR246.

[23] Walsh P, Cunningham P, Rothenberg S J, et al. An artificial neural network ensemble to predict disposition and length of stay in children presenting with bronchiolitis.[J]. uropean Journal of Emergency Medicine, 2004, 11(5):259-264.

[24] Geraci J M, Johnson M L, Gordon H S, et al. Mortality after cardiac bypass surgery: prediction from administrative versus clinical data.[J]. Medical Care, 2005, 43(2):149-158.