Prediction on Potential Health Risks with Regular Physical Examination Record using Machine Learning

Latha A[1], Sunny Kumar Sharma[2], Rahul Kumar Verma[3] , Ribanshu Keshri[4] , Tanmay Gupta[5]

[1]Assistant Professor, Department of Computer Science and Engineering , Sapthagiri College of Engineering, Bangalore, India

[2,3,4,5]Student, Department of Computer Science and Engineering , Sapthagiri College of Engineering, Bangalore, India

---

Abstract— *Today, with modern lifestyle people are now paying more attention towards their well-being. Hence as being self-independent a personalized healthcare is a rising demand among all of us. Now-a-days it can be seen that there is a wide shortage of experienced and knowledgeable personnel like doctors and physicians in almost all healthcare organizations, hence they cannot fulfill the public medical demands. This can be brought to a feasible solution i.e, with the use of hospital data systems, we know there is loads of data generated which is simply getting stacked in databases and may be not used properly. So application can be developed to process such data and to give a customized healthcare service to the public. Here the system proposed is an artificial intelligence system that is AI-assisted prediction system which can predict using data mining methods to establish links between the continuous physical examination of health with its reports and also the potential risks. This system is capable of predicting the subjects risk of physical status which it may have in the coming year based on the reports and observations of current year. The developed system is friendly with users to operate and also facilitates enhanced interface between the entities like doctors and patients. Patients can easily gather their potential health risks which they may have in near future, meanwhile doctors can retrieve patients with their potential risks. This is a feasible way to tackle the incapability to leverage the experience of a good medical healthcare with proper resources.*

***Keywords*—AI-assisted prediction system; Data mining; Potential risks; Physical Examination; Healthcare**

1. INTRODUCTION

Around the world we can witness many healthcare institutions imparting their good services to people. Talking about our neighbouring country china, having immense population and is managing to serve its people in an innovative way in healthcare services. With evolution of technology which is making the whole thing automated and bringing the services to our finger tips, has influenced people to look for their health with same vision with a personalized touch. This demand is also escalated as there is a shortage of trained and experienced medical persons in most of the institutions. This scarcity of providing quality service to people needs to be addressed in a modern way with help of technology. Since there is a lot more than just numbers and figures in the medical reports generated. The data generated everyday in these hospitals are a rich source of hidden and untouched database. Using appropriate data mining techniques and predictive analysis there is a provision to extract this valuable information which can reveal many study patterns. These can be used further to study and advanced analysis on a wide scale set of knowledge for informative and predictive purposes. Tons of datasets are capable to be processed to put them to a suitable and productive uses and extracting links between various entities. As the requirement of researchers to work on the data extracted from various institutions are limited to couple or so, there is a large amount of hidden facts which may not be suitable for direct usage and application. Further these data can be processed to gain knowledge and facts hidden to support a self-service system and enhancing the way of healthcare. This reduces the load of imparting skilled services on clinics which can now be leveraged at fingertips and in less time. Additionally the system is also equipped for use of prognosis or treatment planning, diagnosis.

2. BACKGROUND

In [1], In this paper a deep learning based mechanism is used to research the healthcare data to detect the possible anomalies and classify the info into different in order that we will know the character of ill health. In this paper Deep convolutional neural network (DCNN) is the algorithm that is used to implemented to classify the image patterns data extracted from electrocardiograph (ECG) is discussed intimately. A fanatical convolutional neural network will be trained using different data samples taken from various patients termed as training data. In the later stages, the algorithm is tested using test data samples and it's observed that the algorithm that is used perform efficiently, is stable and is superior classification performance for the detection of normal beats (N-Type), ventricular ectopic beats (V-Type) and super ventricular ectopic beats (SV-Type). The experimental analysis shows the popularity accuracy and loss value. Subsequently, sensitivity and specificity of the algorithm is measured to

point out the effectiveness of the proposed solution. The demerit of this paper is that not having comparison with the fellow studies in this version of results.

In [2], To achieve a specific glucose level the patients with type 2 diabetes are under continuous medical treatment based on anti-diabetic drug. Thus, a sequence of multiple records for prescriptions and their efficacies are related to each patient. These records are embedded by sequential dependencies as personal factors so as that previous records affect the efficacy of this prescription for each patient. During this study, we have utilized the sequential dependencies to render a customized prediction of the prescription efficacy in order to present a patient-level sequential modeling approach .The Recurrent neural networks is used for implementing prediction model that use the sequence of all the previous records as inputs to predict the prescription efficacy at the time this prescription is provided for each patient. Using this approach, each patient's previous records are effectively incorporated for the prediction. The results of experiment of both the regression and classification analyses for the patients having multiple previous records on the real-world data demonstrate improved prediction accuracy.
The disadvantage of this paper is new types of drugs that did not appear in the past data cannot be predicted accurately using the models.

In [3], In the last decade machine learning has gained tremendous interest because of its cheaper computing power and inexpensive memory which makes its efficient to store, process and to analyze high volume of data. To help discover hidden insights and correlations amongst data elements enhanced algorithms are being designed and applied on large datasets not obvious to human. These insights help businesses take better decisions and optimize key indicators of interest. The growing popularity of machine learning also stems from the very fact that learning algorithms are agnostic to the domain of application. Classification algorithms, as an example, that might be applied to categorize faults in windmill blades can also be used for categorizing TV viewers during a survey. The particular value of machine learning however depends on the power to adapt and apply these algorithms to unravel specific world problems. During this paper we discuss two such applications for interpreting medical data for automated analysis. Our first case study demonstrates the utilization of Bayesian Inference, a paradigm of machine learning, for diagnosing Alzheimer's disease supported cognitive test results and demographic data. Within the second case study we specialize in automated classification of cell images to work out the advancement and severity of carcinoma using artificial neural networks. Although these research are still preliminary, they demonstrate the worth of machine learning techniques in providing quick, efficient and automatic data analysis. Machine learning offers hope with early diagnosis of diseases, help patients in making informed decisions on treatment options and may help in improving overall quality of their lives.

The drawback of this paper is the case studies discussed in this paper are initial results. A lot of work lies ahead to expand the analysis to larger and richer datasets to provide more accurate diagnosis.

In [4], once we have an enormous data assail which we might wish to perform predictive analysis or pattern recognition, machine learning is that the thanks to go. Machine Learning (ML) is that the fastest rising arena in computing, and health informatics is of utmost challenge. Machine Learning carries the aim to develop algorithms which may learn and progress over time and may be used for predictions. Machine Learning practices are widely utilized in various fields and primarily health care industry has been benefitted tons through machine learning prediction techniques. The advantage of using machine learning is that it offers a spread of alerting and risk management decision support tools, targeted at improving patient's safety and healthcare quality. With the necessity to scale back healthcare costs and therefore the movement towards personalized healthcare, the healthcare industry faces challenges within the essential areas like, electronic record management, data integration, and computer aided diagnoses and disease predictions. Machine Learning offers a good range of tools, techniques, and frameworks to deal with these challenges. This paper displays the study on various prediction techniques and tools for Machine Learning in practice. A glimpse on the applications of Machine Learning in various domains also are discussed here by highlighting on its prominence role in health care industry.

The disadvantage with the paper is dealing with the huge amount of heterogeneous datasets and increasingly large amount of unstructured and non-standardized information.

In [5], Monitoring health of a person(s) has become a serious IoT application where we'd like to possess a system reception that helps inhabitants to possess their checkup avoided affecting their daily routines. It's considerably essential to possess an IoT driven remote health monitoring system for ailing individuals which may inform caretakers just in case of an emergency. These applications are constrained by the quantity of knowledge collected, managed and exchanged. The factors that constraint IoT applications are robustness, privacy, security and reliability. During this paper, with the help of Internet of Things(IOT) a survey of various health care solutions are presented. This paper presents various challenges and problems that occur in smart health system and proposes solution to beat them.

IOT based solutions suffers from two major challenges:

1) How to secure sensors data.
2) How to provide efficient local and global communication on various devices.

3. MODULE SPECIFICATION

A. Data Cleaning and Transformation

The original dataset is collected from hospital data system. As we are predicting the health risk of an examinee according to his previous health records, we will consider only those examinee who has taken the health examination both the year. These examinees are provided by ID which helps us to select the examinees. After the filtering process we are left out with 2,637 examinees whose data are used. A correct mapping is defined as O : E →{ 0,1}, which may be used to classify critical discriminated of people at risk. We define a mapping O consistent with actual work needs after proper survey of the healthcare centres.

B. Dimensionality Reduction

A high-dimensional feature space can bring many problems to machine learning. Firstly, it significantly increases operating time of prediction. Secondly, because the dimension of features increases, the likelihood of over-fitting increases. Thirdly, the more the experimental variable dimension, the more sparse the info is distributed across the input space, and therefore the harder it's to get a stratified sample of the whole input space. So it's important to use dimensionality reduction methods before data processing process. We use logistic regression to predict beforehand and use the L1 regularization term to guage and choose features. L1 regularization can sparse input matrix, and it helps perform feature selection in sparse feature spaces. We elect 0.1 as threshold of L1, and there are 27 to 30 features remaining after dimensionality reduction varying with different tasks.

C. Classification

Now we've prepared the dataset and ready for applying machine learning methods. This is often core of the whole system, which provides the classification result from feature vector of examinee. We select several different algorithms and compare their performance in our experiments.

• Decision Tree

A decision tree is usually a support tool that resembles the possible consequences of the graphs and model of selections or tree-like graph, which includes the accident outcomes or resource costs and utility. It's a thanks to display an algorithm that only contains conditional control statements. Decision trees are commonly utilized in research, specifically in decision analysis, to assist identify how presumably to realize a goal, but also are a well-liked tool in machine learning.

• XGBoost

This is an optimized distributed gradient boosting library that's designed to be highly efficient, flexible and portable. XGBoost provides a parallel tree boosting (also mentioned as GBDT, GBM) that solves many data science problems during a quick and accurate way. A machine learning technique, Gradient boosting could even be used for regression and classification problems, producing prediction models which are a group of weak prediction models like the decision trees. Like other boosting methods the model is build state wise and allows the optimization of an arbitrary differentiable loss function.

• Random Forest

Random Forests is a learning methods which is used to perform tasks like classification or regression and some others tasks that constructs decision trees in their training time and individual trees of mode of the classes of classification or regression is produced as the output. The overfitting problem of decision tree can be overcome using Random decision forests.

The flowchart of the system depicts the flow of events taking place in every step of the implementation of the system. The feature extraction and applying the algorithm is the key process of the entire system. This helps the system to accurately predict the disease and produce the report.

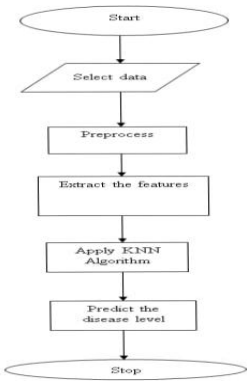

Fig I: Flowchart of the System

4. METHODOLOGY

D. Architecture

First datasets are collected consisting of different attributes such as Hair fall, sickness, arm weakness, blood clots, fever etc. After this we clean these data for any incorrect data formats or some missing data or errors. Then the data is sent for transformation and then for dimension reduction the factors are made redundant free which results in increase in storage capacity.

This data, using the machine learning algorithms are trained again and again for several rounds and then is sent to the client. The client enters the user input or the symptoms and then the client sends it to the trained model

and then the predicted result appear which hare the diseases that the client may have and hence shows prescribed medicines and yoga postures to cure this disease. These helps us to build the self-reliant prediction system which can detect diseases base on the symptoms entered by the client and hence can be used by each and everybody.

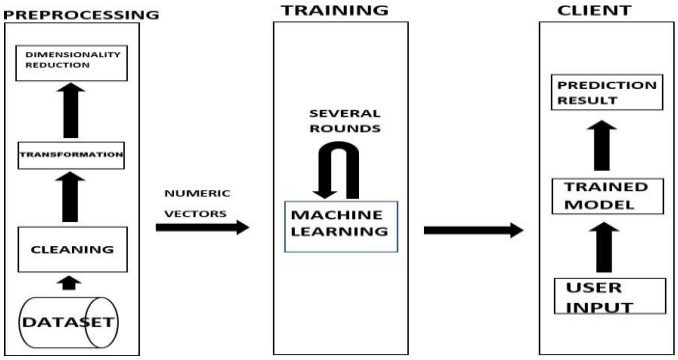

Fig II: System architecture block diagram

E. Algorithms

Naive Bayes: This theorem is a mathematical theorem to detect he output based on certain events. In naïve bayes classifier one cannot see the entire data all at once and hence these data are not inter-related. This also makes them independent from one another. Bayes's theorem is based on posterior probability as

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Here P(A | B) refers to the probability that A will be true if B is true
Also P(B | A) refers to the probability that B will be true if A is true and
P(B) and P(A) refers to the probabilities that B and A will be true respectively.

$$\frac{P(L1 \mid features)}{P(L2 \mid features)} = \frac{P(features \mid L1)P(L1)}{P(features \mid L2)\ P(L2)}$$

The naïve Bayes classifier can be classified into two:
1) Gaussian Naive Bayes
2) Multinomial Naive Bayes

This makes naïve Bayes to have several advantages:
1) They are extremely fast
2) provide straightforward probabilistic prediction
3) Works on small scale data.
4) Highly scalable.

These advantages make naïve Bayes classifier a good choice for classification process. Some of the use cases of naïve Bayes in Machine learning are:
1) Categorizing News
2) Email Spam Detection
3) Face Recognition
4) Medical Diagnosis
5) Weather prediction.

KNN Classifier: k-Nearest Neighbors also called as kNN is a non-parametric and lazy learning algorithm. Some of the techniques to measure distances are:

1) Euclidean distance

$$d_{euclidean} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

2) Manhattan distance

$$d_{manhattan} = \sum_{i=1}^{n}|x_i - y_i|$$

3) By default kNN uses Minkowski distance where y_i and x_i are observations between which the distance is calculated.

$$d_{minkowski} = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p}$$

In machine learning we are trying to find a balance between bias and variance, and depends on the value of k as if :
1) k = n ,where there is high bias but low variance
2) k = 1, where there is low bias but high variance

where n is the number of observations.
An observers class is usually predicted using the classes of its k neighbors. Some of the advantages of kNN are:
1) Simple
2) Executes quickly
3) Accurate
4) No prior knowledge of data.

Since kNN is Simple makes it one of the most used classifiers in machine learning. Some of the use cases of kNN are:
1) Document Classification
2) Delivery store optimization
3) Identifying crime localities
4) Cyber profiling

6. RESULT

The three lab tests includes
1) low density lipoprotein cholesterol LDL-C
2) uric acid or UA
3) triglyceride or TG,

which are usually the key in detection of chronic diseases. The above table consists of
1) Tasks
2) algorithms namely Decision Tree,XGBOOST,RF and Hybrid(Naïve bayes +kNN)
3) Precision
4) recall and
5) F1 score.

The performance scores of precisions, Recall and F1 score are compared for all the four algorithms. F1 score of the four algorithms including the hybrid algorithm i.e. DT, XGBOOST and RF is above 0.75 and hence choosing Machine Learning for detection of diseases what correct

option. XGBoost has performance which stands above the rest of the two algorithms although DT and RF had performances better in some cases like DT has high F1 score in LDL-C and RF has high recall rate in TG etc but comparing it with the Hybrid algorithm that consists of naïve Bayes and kNN classifier, hybrid algorithm has high precision, recall and F1 score hence can be chosen as the main algorithm which is used in this paper and is used to detect the symptoms which are given as input by the client or the user. There is also medications provided based on the symptoms and the diseases as pills and yoga postures to improve the blood circulation and hence cure the disease.

TASK	ALGORITHM	PRECISION	RECALL	F1 SCORE
LDL-C	• DT	0.7997	0.8128	0.8022
	• XGBOOST	0.7997	0.8146	0.7944
	• RF	0.866	0.8039	0.7747
	• NAÏVE BAYES+KNN CLASSIFIER	0.8916	0.9041	0.9077
UA	• DT	0.8426	0.8538	0.8463
	• XGBOOST	0.8498	0.8610	0.8530
	• RF	0.8395	0.8538	0.8429
	• NAÏVE BAYES+KNN CLASSIFIER	0.9256	0.8932	0.8789
TG	• DT	0.7779	0.7611	0.7673
	• XGBOOST	0.7803	0.7807	0.7805
	• RF	0.7697	0.7825	0.7718
	• NAÏVE BAYES+KNN CLASSIFIER	0.9054	0.8891	0.9299

Table I: Performance Scores

7. CONCLUSION

The System provide the potential health risks and warns the user to take preventive measures. The user can give the symptoms and get the predicted health risks. The result is then passed on to the doctor to provide proper guidance. In this busy world people who are not able to reach doctor or do not take minor symptoms seriously can use this system and get their health report. The systems do not need much maintenance as it is a self-service machine. This can also come up as a solution to insufficient and less experienced doctors. The outnumbered doctor and high demand of healthcare can be covered up using this system. As it is a self-service system, it takes up more training data itself and automatically increases the precision. And with the help of doctor we can provide proper precaution and practices like yoga so that people can lead healthy life.

8. REFERENCES

[1] Muhammad Irfan, Ibrahim A. Hameed, Deep Learning based Classification for Healthcare Data Analysis System, 978-1-5386-2366-4/17 ©2017 IEEE.
[2] Seokho Kang Personalized prediction of drug efficacy for diabetes treatment via patient-level sequential modeling with neural networks, https://doi.org/10.1016/j.artmed.2018.02.004 0933-3657.
[3] Niharika G. Maity, Dr. Sreerupa Das, AlaRaj Machine Learning for Improved Diagnosis and Prognosis in Healthcare, 978-1-5090-1613-6/17 ©2017 IEEE.
[4] B. Nithya, Dr. V. Ilango, Predictive Analytics in Health Care Using Machine Learning Tools and Techniques[J], International Conference on Intelligent Computing and Control Systems ICICCS 2017, 978-1-5386-2745-7/17.
[5] Yasmeen Shaikh, V K Parvati, S R Biradar, Survey of Smart Healthcare Systems using Internet of Things(IoT)[J]. 978-1-5386-2459-3/18.
[6] 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.
[7] Delen, D., & Demirkan, H. Data, information and analytics as services. Decision Support Systems, 2013: 55, 359363.
[8] 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.