Perspectives in Communication, Embedded-Systems and Signal-Processing (PiCES)
ISSN: 2566-932X, Vol. 2, Issue 9, December 2018
Proceedings of Knowledge Discovery in Information Technology and Communication Engineering (KITE-2018), 4th May 2018

Lung Disease Prediction over Big Data from Healthcare Communities

Chandana H S[1], Neha Kumari[1], Pallavi R[1], V Padmashree[1], Jayashree M[2]

1. Student, Dept. of Information Science and Engineering, East Point College of Engineering and Technology, Bangalore, Karnataka, India

2. Associate Professor, Dept. of Information Science and Engineering, East Point College of Engineering and Technology, Bangalore, Karnataka, India

Abstract: ***With huge information development in biomedical and social insurance groups, precise investigation of therapeutic information benefits early ailment location, persistent care, and group administrations. Be that as it may, the investigation precision is decreased when the nature of restorative information is fragmented. In addition, distinctive areas display one of kind qualities of certain local ailments, which may debilitate the expectation of ailment episodes. In this paper, we streamline the calculations for viable expectation of lung ailment episode in ailment visit communities. We probe a provincial interminable infection of lung. We propose another convolution neural network (CNN) based multimodal ailment chance forecast algorithm utilizing organized and unstructured information from doctor's facility. To the best of our insight, none of the current work concentrated on the two information writes in the zone of restorative enormous information analytics. Compared with a few common forecast algorithms, the expectation exactness of our proposed calculation achieves 94.8%with a joining speed, which is speedier than that of the CNN-based unimodal ailment chance expectation calculation.***

Keywords: Big data analytics, Healthcare, Lung disease prediction, Hadoop-Map Reduce.

I. INTRODUCTION

The characterization of huge information is turning into a fundamental assignment in a wide assortment of fields, for example biomedicine, online networking, promoting and so on. The current advances in information assembling in a large number of these fields have brought about an unyielding augmentation of the information that we need to oversee. As per a report by the World Health Organization (WHO), the passing because of lung ailments in India were on the ascent representing 11% of the aggregate passing. Upwards of 142.09 in each one lakh, passed on one type of lung malady or the other giving India the questionable qualification of positioning first in lung sickness passing on the planet. Respiratory sicknesses were never again limited to the elderly however were presently being recognized even in more youthful age gatherings. Patients measurable data, test results and sickness history are recorded in the EHR, empowering us to distinguish potential information driven answers for diminish the expenses of medicinal contextual analyses.

Be that as it may, at first the informational collection for patients and their maladies were little with particular conditions which were anticipated by involvement. Any further changes to that pre-indicated informational collection won't give proper disease. With the advancement of enormous information examination innovation, more consideration has been paid to ailment forecast from the point of view of huge information analysis, various investigates have been led by choosing the qualities consequently from an extensive number of information to enhance the exactness of hazard classification as opposed to the already chose characteristics. Risk classification in light of huge information investigation, the accompanying difficulties remain: How should the main chronic diseases in a certain region and the main characteristics of the disease in the region be determined? How can huge information examination innovation be utilized to dissect the malady and make a superior model? To take care of these issues, we join the organized and unstructured information in human services field to evaluate the danger of disease. For unstructured content information, we select the highlights consequently utilizing CNN algorithm.

II. RELATED WORK

Chen et al. [8] proposed a healthcare system using smart clothing for sustainable health checking. Qiu et al. [9] had altogether contemplated the heterogeneous frameworks and accomplished the best outcomes for cost minimization on tree and straightforward way cases for heterogeneous frameworks. Patients' factual data, test results and illness history are recorded in the EHR, empowering us to recognize potential information driven answers for decrease the expenses of therapeutic contextual investigations.

Wang et al. [6] proposed an efficient flow evaluating calculation for the telehealth cloud framework and outlined an information rationality convention for the PHR (Personal Health Record) based dispersed framework. Bates et al. [1] proposed six uses of huge information in the field of human services. Qiu et al. [9] proposed an ideal enormous information sharing calculation to deal with the muddle informational collection in telehealth with cloud systems. One of the applications is to recognize high-hazard patients which can be used to lessen therapeutic

Conference held at Brindavan College of Engineering, Yelahanka, Bagalur Main Rd, Dwarka Nagar, Bengaluru - 560063

cost since high-chance patients frequently require costly human services. Forecast utilizing customary infection chance models typically includes a machine learning calculation (e.g., strategic relapse and relapse investigation, and so on.), and particularly an administered learning calculation by the utilization of preparing information with names to prepare the model. In the test set, patients can be classified into gatherings of either high-hazard or generally safe. These models are important in clinical circumstances and are generally contemplated.

III. PROBLEM STATEMENT

The social insurance groups/biomedical can't anticipate the sickness in view of the patients tried information depictions. There is no exact investigation of medicinal information benefits early malady recognition, persistent care, and group administrations.

IV. OBJECTIVES

- ✓ To predict lung disease using big data from healthcare communities.
- ✓ Less time consumption.
- ✓ Predict the disease at early stage.
- ✓ To improve risk accuracy of risk classification.

V. PROPOSED SYSTEM

This paper plans to build up a multimodal illness forecast utilizing CNN. The real commitments of our work are outlined as takes after:

An expectation framework on a Hadoop stage is worked to foresee the lung ailment which contains three key modules of Information Extraction, Feature Selection and Prediction Modelling.

In Fig.1, Information Extraction is the process where patients record is gathered from doctor's facility and sorted out as EHR (Electronic Healthcare Record).In Feature Selection the highlights of lung ailment expectation and demonstrating is removed from EHR. In Prediction Modelling the given Datasets are analysed to predict whether the patient will get lung malady or not.

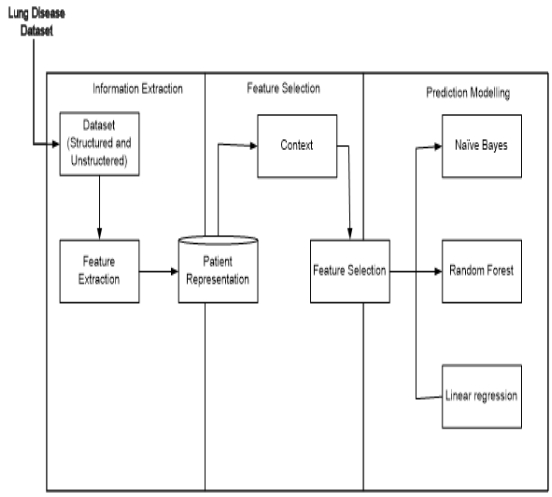

Fig 1. Lung Disease prediction Framework

In Fig.2, the Datasets are the records of patients who are experiencing different lung illness which will be gathered from the hospital. From these Datasets we will remove the normal highlights of various lung diseases, and then this dataset is put away in EHR which speaks to the patients details. The Feature selection makes these records to be put away in an appropriate manner. After doing the prediction, if the symptoms for the particular lung disease are present, then the patient is abnormal or else normal.

Also, advance these information is given to foresee the best possible lung malady by utilizing Naives Bayesian, Random Forest, Linear regression algorithm which gives the precision and decides the stages of the infection.

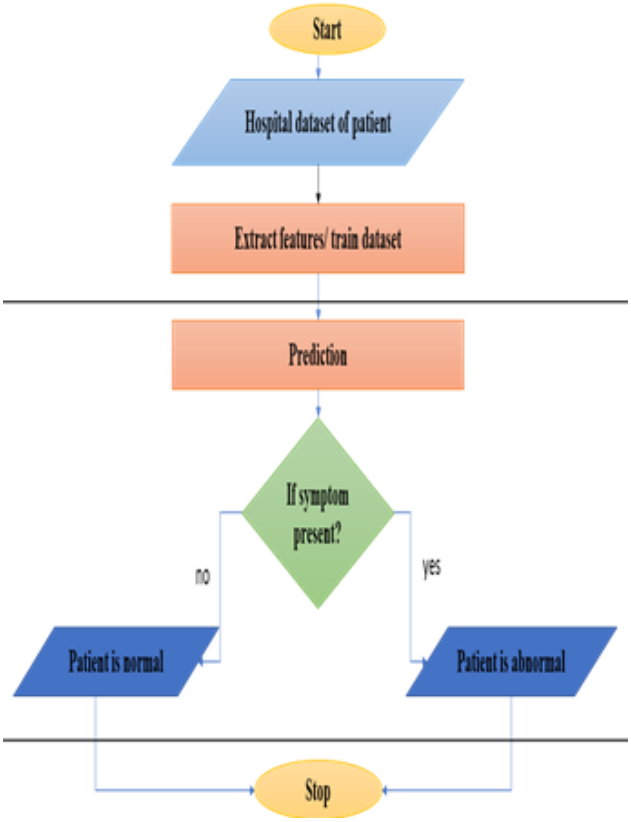

Fig 2. Work Flow of the System

V. PROPOSED SOLUTION

In this section we introduce some background information about the main components used in this project. We propose Naive Bayesian, Random Forest and Linear Regression.

A. NAIVE BAYESIAN

It is a direct and intense algorithm for the classification task. Even on the off chance that we are taking a shot at a data set with many records with some attributes. It takes shot at conditional probability. Conditional probability is

the probability that something will happen, given that something unique has just been occurred. Using the Conditional probability we can figure the probability of an situation utilizing its earlier learning.

Below is the formula for calculating the conditional probability.

$$P(A \mid B) = P(B \mid A) * P(A)/P(B)$$

Where,

- P(A) is the probability of hypothesis A being true. This is known as the prior probability.
- P(B) is the probability of the evidence(regardless of the hypothesis).
- P(B|A) is the probability of the evidence given that hypothesis is true.
- P(A|B) is the probability of the hypothesis given that the evidence is there.

B. *RANDOM FOREST*

Random Forest algorithm is a managed classification algorithm. We can see it from its name, which is to make a forest by some way and make it random. There is an immediate connection between the quantity of trees in the forest and the outcomes it can get: the bigger the quantity of trees, the more precise the outcome. In any case, one thing to note is that making the forest isn't the same as building the decision with information gain or gain index approach.

How Random Forest algorithm works?

There are two stages in Random Forest algorithm, one is random forest creation, the other is to make a prediction from the random forest classier created in the first stage. The whole process is shown below,

1. Randomly select “K” features from total “m” features where k <<m.
2. Among the “K” features, calculate the node “d” using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat the 1 to 3 steps until “l” number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

In the next stage, with the random forest classier created, we will make the prediction. The random forest prediction steps are shown below:

1. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the votes for each predicted target.
3. Consider the high voted predicted target as the final prediction from the random forest algorithm.

C. *LINEAR REGRESSION*

Linear regression is an extremely basic technique however has turned out to be exceptionally helpful for countless.

Below is the raw data.

1	x	y
2	1	1
3	2	3
4	4	3
5	3	2
6	5	5

The attribute x is the symptom variable and y is the disease predicted variable that we are trying to predict. If we got more data, we would only have x values and we would be interested in predicting y values.

VI. RESULT

In the Graph the results displayed based on the accuracy, the comparison of accuracy we can see with the Random Forest, Linear Regression and Naive Bayesian classifiers. Here Random Forest achieves least accuracy and Naive Bayesian achieves more accuracy. The dataset we are taken with major symptoms for the lung cancer disease.

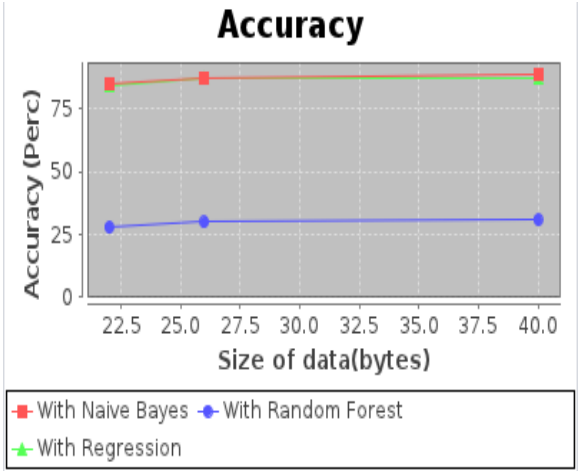

Fig 3. Measure of accuracy

Here we can say 9 major attributes as symptoms. In the dataset each line tells the details of one patient with different symptoms as attributes. We need to take more patient dataset to predict the disease.

VII. CONCLUSION

We focused on the different lung malady forecast utilizing organized and unstructured dataset and hence proposed a Multi-model CNN utilizing Naive Bayesian, Random Forest and Linear Regression on Hadoop platform, to decrease time utilization, foresee the sickness at beginning period and to enhance predicting the

precision of risk classification. It is made out of information extraction module, feature selection module and prediction modelling. Furthermore, we assess the exactness of the ailment being anticipated by the algorithm.

REFERENCES

[1] D. W. Bates, S. Saria,L.Ohno-Machado, A. Shah, and G. Escobar, ''Big data in health care: Using analytics to identify and manage high-risk and high-cost patients,'' Health Affairs, vol. 33, no. 7, pp. 1123–1131, 2014.

[2] D. Oliver, F. Daly, F. C. Martin, and M. E. McMurdo, ''Risk factors and risk assessment tools for falls in hospital in-patients: A systematic review,'' Age Ageing, vol. 33, no. 2, pp. 122–130, 2004.

[3] S.Marcoon, A. M. Chang, B. Lee, R. Salhi, and J. E. Hollander, ''Heart score to further risk stratify patients with low TIMI scores,'' Critical Pathways Cardiol., vol. 12, no. 1, pp. 1–5, 2013.

[4] S.Bandyopadhyay et al., ''Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data,'' Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1033–1069, 2015.

[5] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, ''A relative similarity based method for interactive patient risk prediction,'' Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.

[6] J. Wang, M.Qiu, and B. Guo, ''Enabling real-time information service on telehealth system over cloud-based big data platform,'' J. Syst. Archit., vol. 72, pp. 69–79, Jan. 2017.

[7] J. Wan et al., ''A manufacturing big data solution for active preventive maintenance,'' IEEE Trans. Ind. Informat., to be published, doi: 10.1109/TII.2017.2670505.

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

[9] M. Qiu and E.H.-M. Sha, ''Cost minimization while satisfying hard/soft timing constraints for heterogeneous embedded systems,'' ACM Trans. Design Autom. Electron. Syst., vol. 14, no. 2, p. 25, 2009.