International Engineering Journal For Research & Development
Vol.5
Issue 5
E-ISSN NO:-2349-0721
Impact factor : 6.549

DETECTION OF DISEASES OVER GEOGRAPHICALLY DIVERSE LOCATIONS USING MEDICAL SYMPTOMATIC DATA AND NATURAL LANGUAGE PROCESSING

[1]Shweta Suresh Mahavarkar, [2]Dr. A. N. Cheeran, [3]Shweta Yadav
Electrical Department Veermata Jijabai Technological Institute Matunga, India[1,2]
, R & D Engineer A3 Remote Monitoring Technologies Pvt Ltd India[3]
mahavarkarshweta@gmail.com[1], ancheeran@ee.vjti.ac.in[2], shweta.yadav@a3rmt.com[3]

ABSTRACT

Early detection of regional symptomatic diseases can lead to a decreased impact on populations. Without having clear information about necessary conditions for earlier detection and their influencing factors, attempts to improve surveillance will be unsystematic. Systematic methods can be developed by considering large symptomatic data and natural language processing. Those symptomatic data is available in special comments made by paramedic/doctor during physiological data acquisition from A3 Critiview. We can access those special comments and location of it from SQL database and A3 cloud server respectively. Natural language processing on statistical data analysis can find out common diseases over particular locations. The software used for designing the proposed system is Python 3.

Keywords—regional symptomatic disease; natural language processing; physiological data acquisition; statistical data analysis

INTRODUCTION

A disease is any condition which causes malfunctioning of the body. Diseases are recognized by a specific set of symptoms such as a cold, the flu, measles, cancer, stroke, or diabetes, just to name a few. Specific diseases such as communicable diseases spread from person to person or from animal to person. Transmission or spread can happen through air, water, contact or through some other medium. Disorders that are caused by organisms - such as bacteria, viruses, fungi or parasites are termed as infectious diseases. Signs and symptoms of any disease vary depending on the type of disease. These diseases keep on spreading over different locations which lead to outbreak / epidemic / pandemic. Epidemics and pandemics have always had major impacts on the affected populations both socially and economically. For example, 2009 flu pandemic in India is the outbreak of swine flu in various parts of India. There were around 1400 death of individuals in 2010 in India because of swine flu [9]. Also, the latest outbreak of Coronavirus disease has affected worldwide. More than 7000 deaths have been reported till now and it is getting multiplied at a very faster rate [9]. Hence, detecting the spread of such epidemics / pandemics at an early stage across various locations will be helpful for early diagnosis and for death prevention on a larger scale. After identification of an emerging pandemic, detecting the disease spread, local and international healthcare organizations can be notified earlier so that they can take steps to halt the disease's progress. Thus, controlling the epidemic diseases at the beginning of its spread is a vital solutions for epidemics [1].

The proposed system aims to develop an artificially intelligent system which can detect the spread of diseases at an early stage over different geographical locations. This can be done by extracting the regional symptomatic data and location from SQL database and the Cloud Server respectively and then perform Natural Language Processing (NLP) on the special comments made by the paramedic/doctor during the physiological data acquisition. Another aim is to identify and detect common diseases over particular geographical regions by statistical data analysis with the help of NLP. The other vital parameters like SPO2, PR, ECG, HR, NIBP, Body temperature can be acquired from the database and the information can be given to the doctor in real time. The patient data will be stored on the web server so that the doctor can access the information whenever required from anywhere and does not have to be physically present.

There are various researches done in the field of healthcare for any disease detection and prediction at an early stage. In [2], Diagnosis of plant disease through a Cloud based scalable collaborative platform is done where farmers / users upload the images of plants for automatic disease diagnosis in real time. The uploaded image gets classified by the AI engine. For knowing the particular disease in the affected location, geo-location of image and time stamp is used. The paper [2] is similar to the proposed work done which does disease diagnosis for plants with the help of image processing. There are many papers which use early disease detection techniques [1][3][4]. In [3], author has applied different classification algorithms, each with its own advantage on three separate databases of disease (Heart, Breast cancer, Diabetes) available in UCI repository for disease prediction. The feature selection for each dataset was accomplished by backward modeling using the p-value test. The results of the study strengthen the idea of the application of machine learning in early detection of diseases. In [1], the author has designed a warning and detection system for epidemic diseases. An algorithm using GPS is designed by the author to prevent disease infection at the source and international propagation. The system is divided into 3 parts: the control center of the epidemic disease, the traveler's smart phone app, and the GPS log analyzer at airports. In [4], author has stated that the clinical symptoms of Diabetes Mellitus (DM) can be used for early detection . The variables used in this study are symptoms and factors supporting DM .The patient data is collected and then mapped into data that is ready to be used in the process of artificial neural networks in the form of neuron input and neuron output. The main aim of the author is to reduce the no of patients having delayed treatment for diabetes. NLP has been used in many papers for medical record classification [5][6][7][8]. In [5], the goal of the author was development of a software tool to detect documentation of Pediatric Appendicitis Score (PAS) within electronic emergency department (ED) notes. PAS should fall within a certain range so as to minimize patients' exposure to radiation. The software application was developed using a combination of NLP and Machine Learning (ML) methods (i.e. Statistical NLP). In [6], the author aims to identify symptoms of a disease formed from one or more words. Names Entity recognition (NER) is the method using NLP to identify symptoms of the digestive disease. In [7], the author has grouped, classified the symptoms and then score-based analysis is done. In [8], the author talks about algorithm for entity extraction, its observation and implementation for Chinese text. The algorithm is divided into five steps: rules analysis, statistic analysis and syntax analysis. Rule analysis is semantic analysis i.e understanding the meaning of words; Statistics analysis is analysis of new words and creating a dictionary for any new word; Syntax analysis is nothing but arrangement of words so as to make grammatical sense. This paper shows the overall idea and implementation of natural language processing. The related work done until now includes early detection of epidemics, some specific disease detection using NLP, used NLP for classifying the electronic health records. The proposed system will predict the spread of the disease over different geographical regions.

The prediction will be done by accessing the symptomatic data whenever needed from anywhere and need not be physically present.

In Section II a brief discussion about the proposed system block diagram will be done. Section III will be about the Result and Discussion. Lastly, Conclusion and Future work will be discussed in Section IV.

PROPOSED SYSTEM DESIGN

In this section, the illustration and flow of the proposed system design will be discussed. Fig. 1 shows the block diagram of the proposed system. The system consists of an Application Program or a Processor which will take three major inputs which are Time at which the patient record was taken, GPS (latitude and longitude) co-ordinates to give information about the location and the Symptom(comments) made by the paramedic / doctor. These inputs can be accessed from the SQL database and the A3 cloud server. Database will consist of the patient details, the vital parameters (SPO2, NIBP, etc.) of the patients and also the symptoms of the patient given by the doctor in the form of comments. The processor will process the data present in the database and will generate different reports by taking into consideration the various parameters. Some of the analytical reports that will be generated are SPO2 analysis report, NIBP analysis report for a particular location, age, gender. An efficient and a user-friendly user interface will be developed which will help for accessing the reports in various forms.

NLP is the ability of the computer for making human interaction with the machines. It includes semantic and syntax analysis. Syntax is arranging words in a sentence to make sense grammatically to the sentence. Semantics involves the use and meaning behind words [8]. NLP includes sentence and word tokenization, stemming of words, segmentation, removing the unnecessary words or the stopwords [6]. The comments or the symptoms made by the paramedic will then be further processed by the processor using NLP. These comments which will be in the form of symptoms will undergo NLP and specific / particular keywords will be identified. Depending on these keywords the disease can be predicted. With the help of NLP we can predict disease over different geographical regions.

E-ISSN NO:2349-0721

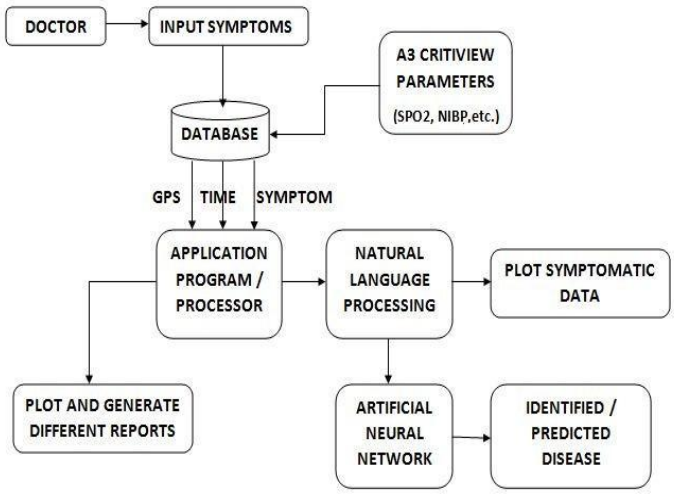

Fig. 1. Proposed System Block Diagram

Different reports can be extracted after performing NLP. For predicting the disease, artificial neural network (ANN) will be used along the NLP. An ANN is based on the neural structure of the brain which can learn to perform tasks like classification, prediction, decision-making, visualization. Finally, the disease can be predicted

by using ANN by natural language processing. The software used for performing the above is Python since it is a high-level, object-oriented programming language which has extensive libraries and toolkits available.

RESULT AND DISCUSSION

In this section, a brief discussion of the results will be done. Fig. 2. shows the user interface developed so as to extract report using the parameters of patients like SPO2 and NIBP. The data can be analyzed depending on the gender of the patient. The report can be generated for males individually, females individually and also for both. This user interface can be used for predicted disease analysis and also gives the geographical view.

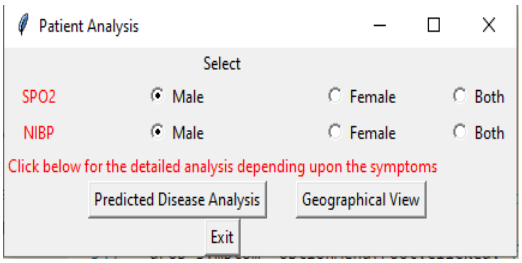

Fig. 2. User Interface for Report Analysis

Fig. 3. and Fig. 4. show the report analysis for Blood Pressure and SPO2 respectively. The plot 'Patient Analysis depending on BP' shows the no of patients whether male or female in different categories like Low BP, Ideal BP, Pre-High BP, High BP. SPO2 plot shows the no of patients both male and female having SPO2 less than 95% which means that the oxygen level in the patient is less than normal. These reports can be used by various health departments for further analysis.

Fig. 3. Patient Analysis depending on NIBP

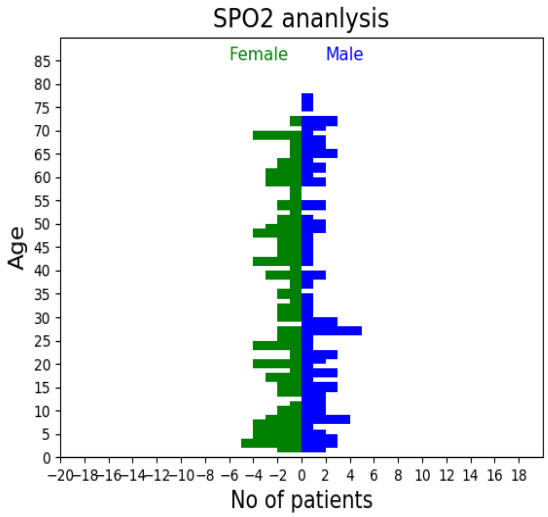

Fig. 4. Patient Analysis depending on SPO2

Fig. 5. shows the Symptomatic Data plot after performing NLP over almost 200 patient records. Symptoms here are made in the form of sentences by the paramedic / doctor. Sentences undergo NLP and specific keywords relating to the symptoms are retained from the sentence. NLP includes tokenization of words, removing the unwanted words and then stemming the words. The keywords identified are viral-fever, eyes, diarrhea, skin, asthama and count has been found for each of them respectively

Fig.6. shows the plot for disease prediction count for specific diseases (considered two - dengue and tuberculosis) based on the symptomatic data after performing NLP. These diseases have some specific set of symptoms based on which the grouping is done. Grouping is done depending on how many symptoms or the no of symptoms the patient has been diagnosed with. The plot shows the count of detected and not detected patients for both dengue and tuberculosis.

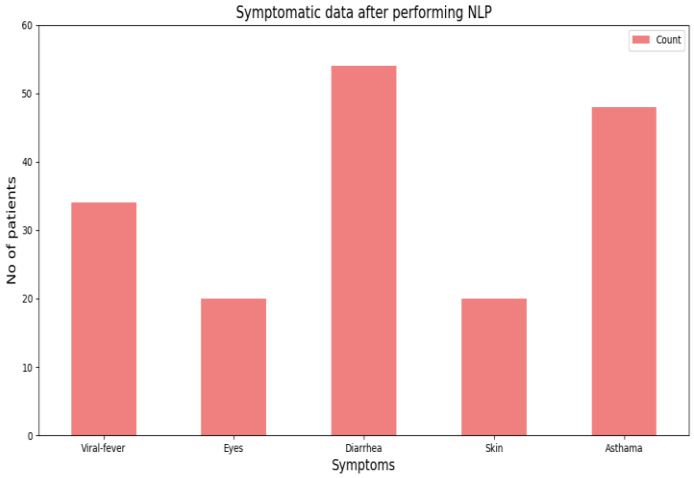

Fig. 5. Symptomatic data plot after performing NLP

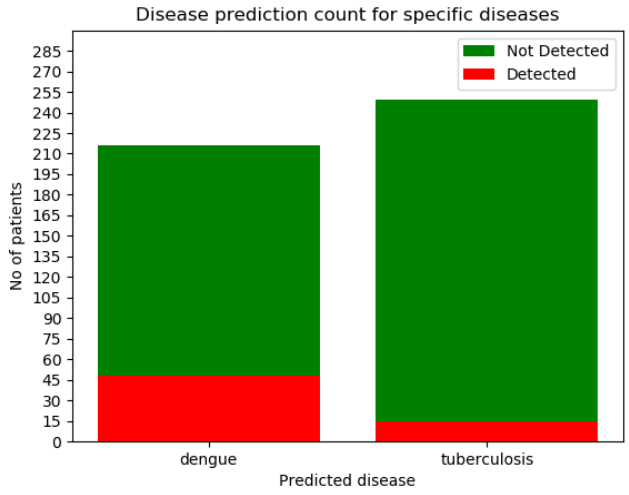

Fig. 6. Plot of Disease prediction count

In order to know the spread of some specific set of symptoms over different regions, the geographical or the map-view is required. This geographical plot will show the presence of any symptom in the particular region. Fig. 7. shows the User Interface created for plotting symptomatic data after performing NLP on a geographical map.

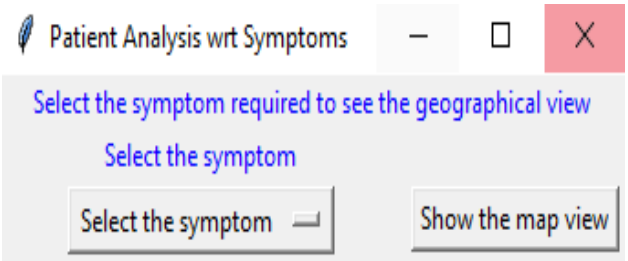

Fig. 7. User Interface for Symptom selection

After selecting any particular symptom from the dropdown of the user interface, the map-view will show the no of patient spread across different regions having the selected symptom. Fig. 8. shows the geographical view of the symptom 'cough' in the states of India. This geographical view shows the total count of patients having cough. When the cursor moves over to the coordinates of any particular state, cursor will show the count of the patients in that state as displayed the count of patient for the state Uttar Pradesh in the graphical plot. The dataset used for geographical plotting consists of more than 250 entries.

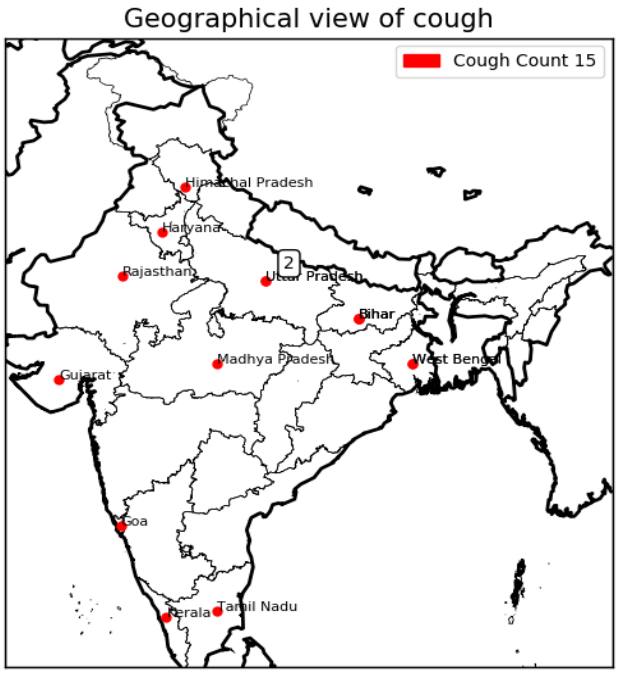

Fig. 8. Geographical view of the selected symtom (cough)

CONCLUSION AND FUTURE WORK

Detection of various diseases will improve the healthcare services. Increasing rate of communicable / infectious diseases has a very high impact on the population. Detecting the epidemics and knowing the spread of the disease is an essential factor to reduce the impact on the population. This proposed system aims at detecting the disease by analyzing different parameters of the patients; and also by analyzing the medical symptomatic data and then applying machine learning algorithm and natural language processing on that data. NLP is adopted or implemented so that large symptomatic medical dataset can be visualized in a smaller version. The reports generated as discussed in the above results can be used by the medical or the health departments for analysis. The geographical map plot can be used to know the spread of any disease across different location which will help early detection and diagnosis. Data Visualization plays a vital role in representing your information and data graphically in enormous ways. It also gives ideas to enhance your work more precisely and efficiently by analyzing, sorting, grouping, mapping the data. The future work is to visualize the data more prominently and add animations to the graphs. These graphs can consider many other attributes like year range, age, gender, etc. The future aim is to perform NLP over Artificial Neural Network which can be used for decision making, classification, prediction of disease over a larger database.

REFERENCES

[1] M. Kim, J. Y. Lee, and H. Kim, "Warning and detection system for epidemic disease," 2016 International Conference on Information and Communication Technology Convergence (ICTC), 2016.

[2] K. K. Singh, "An Artificial Intelligence and Cloud Based Collaborative Platform for Plant Disease Identification, Tracking and Forecasting for Farmers," 2018 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 2018.

[3] P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018.

[4] F. Aofa, P. S. Sasongko, Sutikno, Suhartono, and W. A. Adzani, "Early Detection System Of Diabetes Mellitus Disease Using Artificial Neural Network Backpropagation With Adaptive Learning Rate And Particle Swarm Optimization," 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS), 2018.

[5] B. Norman, T. Davis, S. Quinn, R. Massey, and D. Hirsh, "Automated identification of pediatric appendicitis score in emergency department notes using natural language processing," 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2017.

[6] F. B. Putra, A. A. Yusuf, H. Yulianus, Y. P. Pratama, D. S. Humairra, U. Erifani, D. K. Basuki, S. Sukaridhoto, and R. P. N. Budiarti, "Identification of Symptoms Based on Natural Language Processing (NLP) for Disease Diagnosis Based on International Classification of Diseases and Related Health Problems (ICD-11)," 2019 International Electronics Symposium (IES), 2019..

[7] K. Duangchaemkarn, V. Chaovatut, P. Wiwatanadate, and E. Boonchieng, "Symptom-based data preprocessing for the detection of disease outbreak," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017.

[8] D. Feng, Y. Baozong, and L. Biqin, "Extracting entities for natural language dialog system," WCC 2000 - ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000.

[9] "Disease outbreaks by year", World Health Organization, 2020. [Online]. Available: https://www.who.int/csr/don/archive/year/en/. [Accessed: 18- Mar-2020].