Proposed Model for Predictive Diagnosis of Chest Diseases

[1]Advait Srinivas, [2]Akshay Solunke, [3]Pranjal Naik, [4]Umesh Kulkarni

[1]BE Student, [2]BE Student, [3]BE Student, [4]Assistant Professor
[1]Department of Computer Engineering,
[1]Vidyalankar Institute of Technology, Mumbai, India

Abstract: the proposed system is a predictive analysis tool for medical diagnosis using machine learning for chest conditions like Asthma, COPD, and Tuberculosis. All participants will have a questionnaire (self-administered / physician-administered). Positive and negative predictive values calculated for each question and the total scores of patients compared with those of control. This tool will enable doctors with data to make better and early differentiated diagnosis minimizing the errors and timely treatment.

Index Terms **- Chest Diseases; Tuberculosis; Asthma; COPD; Symptom-Based Questionnaire; Decision Trees; Linear Regression;**

I. Introduction

There are two key problems regarding treatment of chest diseases in human beings globally. First, they are often diagnosed later than they should be due to patient negligence. Second, they are sometimes misdiagnosed when there is no access to skilled medical personnel. Our proposed system aims to solve this problem for prominent chest diseases like Asthma, COPD, Tuberculosis, etc.

The approach to building the system involves using a symptom-based questionnaire and implementing machine learning algorithms to predict from which disease the patient might be suffering. Each chest disease has symptoms that demonstrate its presence. These symptoms include shortness of breath, chest congestion, chest pain, cough from the throat, and cough from the chest, etc. and manifest in different situations when human beings are functioning in their day to day lives. We seek to use these symptoms and their manifestation in different daily cases such as running, laughing, talking, etc. to detect which chest disease the human being might be facing. To make the machine train on the sample datasets containing symptoms in questionnaire form.

This system will enable patients to perform self-diagnosis to approach doctors sooner. It will help nursing homes, medical centers and public hospitals perform an initial diagnosis measure before further check-ups by specialized doctors.

II. Existing System

The previous systems designed for chest disease prediction were developed with the intent of getting a perfect predictive diagnosis. Therefore they were designed to collect large amounts of medical report data from advanced testing like spirometry, pulmonary function test, and other tests. Based on these tests, machines were trained to predict which disease had occurred.

However, these systems are expensive and often require advanced medical equipment and medical professionals to conduct tests, collect data, administer policies, etc. In addition, it is difficult to scale and use by the average user that wants a quick diagnosis or for nursing homes, medical centers, etc. to perform diagnosis where there is no advanced medical personnel available.

III. PROPOSED SYSTEM

Our proposed system aims to create an initial predictive diagnosis that can be scaled and used by anyone. Over the years medical researchers have arrived at a synthesis of this medical data to give us symptom-based questionnaires that can be used by people to detect these diseases. But the limitations of these questionnaires are that they have been arrived at in small clinical trials using small amounts of patient and control data. Therefore there is a need to build a machine learning system that uses large amounts of patient and control data to verify and use these symptom-based questionnaires for the broader public.

We seek to integrate several of these symptom-based questionnaires with real-life scenario data to able to precisely and yet easily predict which chest disease the patient has. There are two kinds of data required, patient data (chest disease patients and their symptoms) and control data of people who don't suffer from these diseases but show some signs of chest diseases. By integrating these data sets, to create weighted scores for each question in the questionnaire, we will be able to generate a result of which chest disease the patient is suffering.

Our system aims at utilizing supervised machine learning algorithms and inbuilt python libraries to train the machine. Training of machine is done using datasets from the UCI database, California Health and Human Services (CHHS) data-portal, US Government Open Data (data.gov) for training and testing on US-based data. Indian-based data is obtained from partnerships with reputed local hospitals, physicians and chest specialists.

IV. STEPS IN BUILDING THE PROPOSED SYSTEM

Our proposed system is designed to be easily used by anyone thanks to the symptom-based questionnaire that allows for simple and seamless user experience with the tool. The user needs to enter the answers to the symptom-based questionnaire and will receive a report. The report contains the following parameters - percentage to indicate the probability of the presence of chest disease and a graph that shows which conditions match and which do not match. Steps involved in building the system:

4.1 Questionnaire Generation

We collected sample questionnaires created by medical researchers in the United States and Europe. [1] Each questionnaire has been made for a different disease such as Asthma, Tuberculosis, Lung Cancer, etc. [4] [5] However, no questionnaire has been applied for several chest diseases at a time. At the same time, these sample questionnaires have also made for Western patients in the US and Europe. [3] Therefore we generated a global questionnaire that integrates symptoms for Asthma, COPD, Tuberculosis, and Bronchiectasis based on the different sample questionnaires for Indian patients.

The global questionnaire contains basic/fundamental symptoms that indicate whether the patient has any chest-related disease or some other condition. These basic/fundamental symptoms include shortness of breath, chest congestion, coughing for long periods, etc. The questionnaire also has differentiated symptoms like a progressive increase in symptoms versus regularly occurring symptoms, sputum production, etc. It also contains secondary symptoms such as headaches, weight loss, etc. that occur along with some chest diseases. It also includes environmental conditions like exposure to smog, dust, etc. and genetic conditions like allergies or a history of lung disease in the family that could have triggered the chest disease.

Most sample questionnaires available contained only one or two dimensions that indicate chest disease. [2] We created a combination of fundamental, differentiated, secondary, environmental and generic symptoms for the global questionnaire that forms the basis of precisely diagnosing which chest disease the patient is suffering.

4.2 Machine Training with Appropriate Data Sets

Each of these questions in the questionnaire has equal weighted scores assigned at the beginning. The machine .is then trained with data from patients who are suffering from the disease and from healthy people who are not suffering from the disease but show a few symptoms indicated.

The nature of the data sets collected includes a significant amount of statistics indicating patient profile details such as smoker or non-smoker, age, occupation, etc. that set the broad indicators of chest disease. But the datasets also include statistics of genetic conditions like allergies, environmental and geographical conditions and even statistics of the nature of fundamental, differentiated and secondary symptoms that patients/healthy individuals face over one year. The weighted score of specific question changes based on the training data (patient and control data). As the weight of the question increases or decreases, it has a higher value in the disease classification process.

Machine Training is done using Supervised Learning Algorithms because the outcome is known. This problem is one that requires a learning algorithm conducive for both classification and regression, and therefore we use the Decision Tree Algorithm. The first goal of the training is to classify the data that affects the weight of the question. [6] The nature of the data collected is in the form of simple yes or no which also facilitates the use of decision trees to architect the learning decision rules inferred from the data. [7] The training is carried out with the help of an API like TensorFlow that has pre-installed Python libraries.

4.3 Data Mining and Linear Regression

Following the training, the questionnaire has adjusted weights. The questionnaire becomes a data mining mechanism to classify if the patient has a chest disease. User enters the data in the questionnaire and the user data is processed and stored in the database. This data is then compared to the data sets to check for matching conditions and non-matching conditions. If conditions do not match, the process continues to check until all the questionnaire parameters are completed. If conditions match, it continues to check for other conditions and carries over the data to the next phase to classify from which chest disease the patient is suffering. To calculate the probability percentage and generate a graph of the newly classified chest disease based on the data entered by the user, we use linear regression algorithm.

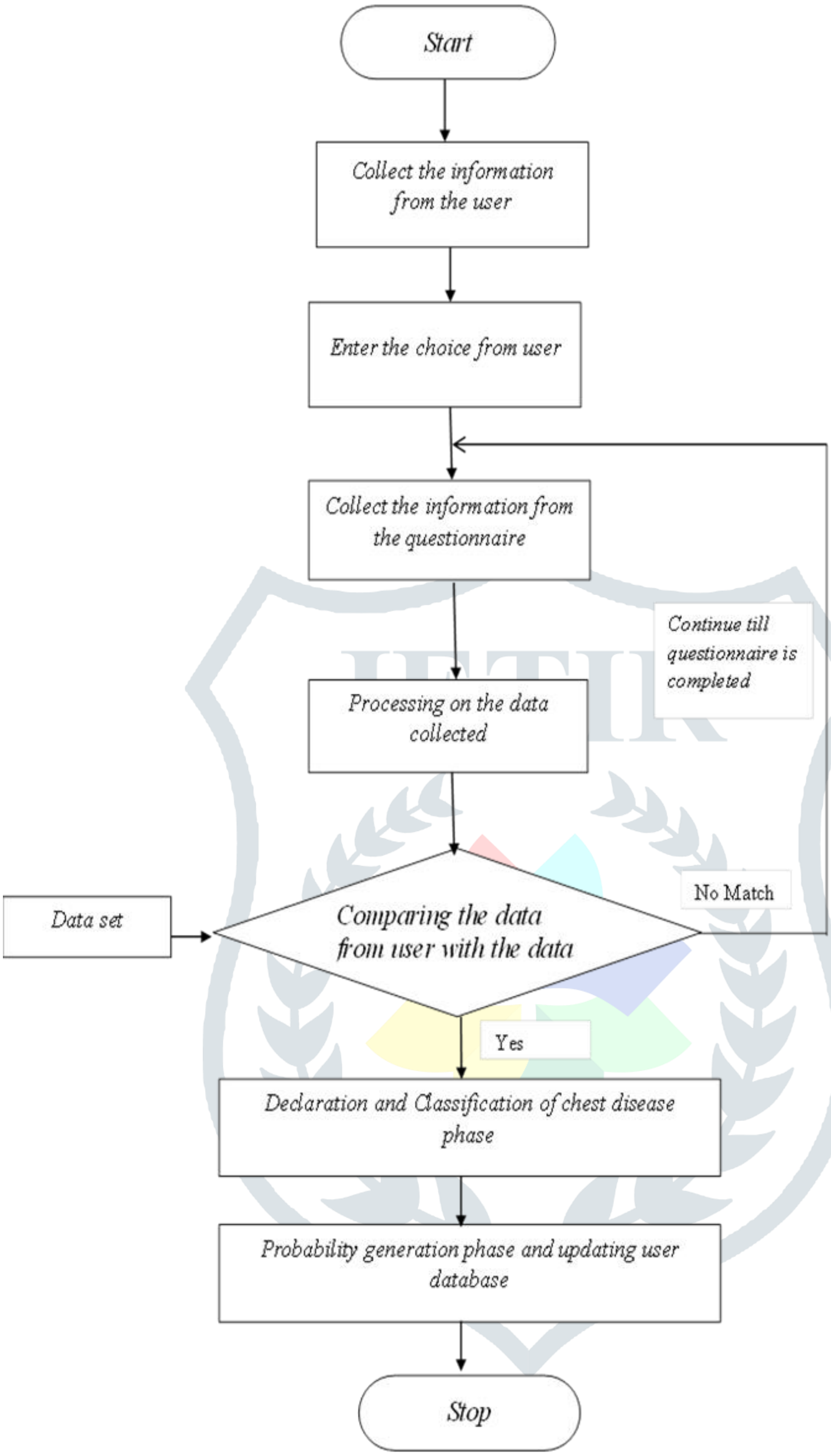
Start
Collect the information from the user
Enter the choice from user
Collect the information from the questionnaire
Continue till questionnaire is completed
Processing on the data collected
Data set
Comparing the data from user with the data
No Match
Yes
Declaration and Classification of chest disease phase
Probability generation phase and updating user database
Stop

Fig 1: Machine Flow Diagram

V. USER EXPERIENCE

The user enters data in three distinctive steps when interacting with the system following which he receives a diagnosis report with which he can take further decisions. The user experience includes the following steps –

5.1 User builds a basic health profile

This step includes attributes such as age, gender, occupation, weight & height resulting in BMI which is an indicator of general fitness. The user also enters factors like whether he/she is a smoker or not. This step creates a basic health profile with critical indicators of competence and potential chest disease.

5.2 User inputs the data based on the questionnaire to create an advanced medical profile

This step is to identify and predict if the user suffers from a chest disease and which chest disease it could be. It includes factors like the daily manifestation of symptoms, environmental factors, family history, allergy, side effects, etc. Each question has a weighted score and associated heuristics which gets fed into the decision tree.

5.3 Rechecking to fix the prediction of the disease

User encounters a few more questions to determine the exactness and precise probability of the condition identified in the second step. The system works on this input to generate a probability percentage of the disease occurring.

5.4 Report Generation

A report is generated based on the inputs that include – the probability of the disease, comparative graphs concerning other patients and potential next steps that the patient can take.

.

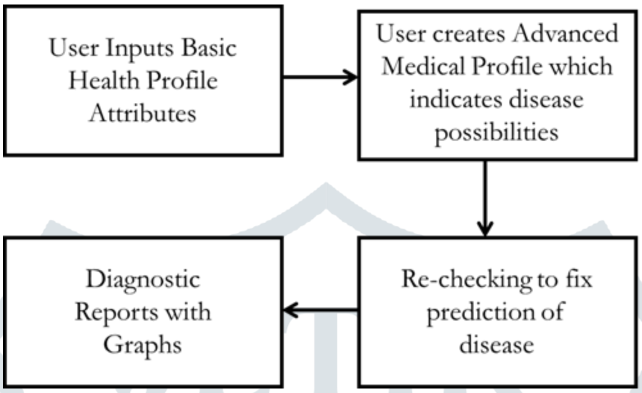
User Inputs Basic Health Profile Attributes
User creates Advanced Medical Profile which indicates disease possibilities
Re-checking to fix prediction of disease
Diagnostic Reports with Graphs

Fig 2: User Experience Flow

VI. Conclusion

The scope of the project is to aid the initial diagnosis of chest diseases and to help in the identifying differentiation between these diseases. Our project uses the concept of symptom-based questionnaires and weighted scores assigned for these questions. We would be designing the plan such that it incorporated into the daily working of any local physician, nursing home or hospital.

The project is not a complete predictor of disease but an indicator for patients and doctors to minimize negligence and to avoid unnecessary delay in taking further action.

VII. Acknowledgment

We want to express our gratitude towards the teaching and non-teaching faculty of Vidyalankar Institute of Technology for the resource and guidance in the completion of this project.

We want to express special gratitude and thanks to the physicians who gave us their valuable time and expertise. We especially thank Dr. Ashish Naik, Medical Officer, Bhandara District for his guidance in developing the global questionnaire and providing useful insights about the state of chest disease treatment in smaller cities and villages in Maharashtra and the rest of India.

References

[1] A New Symptom-Based Questionnaire for predicting the presence of Asthma – B Shin, SL Cole, S-J Park, DK Ledford, RF Lockey Division of Allergy and Clinical Immunology, Department of Internal Medicine, University of South Florida College of Medicine, James A. Haley Veterans'Medical Center, Tampa, Florida

[2] Symptom-Based Questionnaire for Differentiating COPD and Asthma, 2006;73:296-305. doi:10.1159/000090141 - Tinkelman D, G, Price D, B, Nordyke R, J, Halbert R, J, Isonaka S, Nonikov, D, Juniper E, F, Freeman D, Hausen T, Levy M, L, Østrem A, van der Molen T, van Schayck C

[3] Symptom-Based Questionnaire for Identifying COPD in Smokers, Respiration 2006;73:285-295. doi: 10.1159/000090142 - Price D, B, Tinkelman D, G, Halbert R, J, Nordyke R, J, Isonaka S, Nonikov D, Juniper E, F, Freeman D, Hausen T, Levy M, L, Østrem A, van der Molen T, van Schayck C

[4] COPD – Differentiated Diagnosis - GrantHoekzema, MD Program Director, Mercy Family Medicine Residency, St. Louis, MO, ElissaJPalmer, MD, FAAFPProfessor and Chair, Department of Family & Community Medicine, University of Nevada School of Medicine, Las Vegas, NV

[5] http://www.tbcontrollers.org/docs/TBDrugsAndBiologicsShortages/Delaware_TB_Symptom_Screening_Questionnaire

[6] A Tree-based Decision Model to Support Prediction of the Severity of Asthma Exacerbations in Children. (Ken Farion Departments of Pediatrics and Emergency Medicine, University of Ottawa, Ottawa, Canada Wojtek Michalowski, Szymon Wilk1 , Dympna O'Sullivan Telfer School of Management, University of Ottawa Ottawa, Canada Stan Matwin School of Information Technology and Engineering, University of Ottawa Ottawa, Canada Institute of Computer Science, Polish Academy of Sciences Warsaw, Poland)

[7] Development of a diagnostic decision tree for obstructive pulmonary diseases based on real-life data (Esther I. Metting, Johannes C.C.M. in 't Veen, P.N. Richard Dekhuijzen, Ellen van Heijst, Janwillem W.H. Kocks, Jacqueline B. Muilwijk-Kroes, Niels H. Chavannes and Thys van der Molen)