VIVA-Tech International Journal for Research and Innovation
ISSN(Online): 2581-7280

Volume 1, Issue 2 (2019)
Article No. 14
PP 1-4

VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

Proposed Model for Chest Disease Prediction using Data Analytics

Vikrant A. Agaskar [1] and Umesh Kulkarni[2]
[1](PG Student ARMIET, Dist. Thane, India)
[2](Vidyalankar Institute of Technology, India)

Abstract: *Chest diseases if not properly diagnosed in early stages can be fatal. Because of lack of skilled knowledge or experiences of real life practitioners, many a times one chest disease is wrongly diagnosed for the other, which leads to wrong treatment. Due to this the actual disease keeps on growing and become fatal. For example, muscular chest pains can be treated for the heart disease or COPD is treated for Asthma. Early prediction of chest disease is crucial but is not an easy task. Consequently, the computer based prediction system for chest disease may play a significant role as a pre-stage detection to take proper actions with a view to recover from it. However the choice of the proper Data Mining classification method can effectively predict the early stage of the disease for being cured from it. In this paper, the three mostly used classification techniques such as support vector machine (SVM), k-nearest neighbour (KNN) and artificial neural network (ANN) have been studied with a view to evaluating them for chest disease prediction.*

Keywords *– KNN, SVM, Data Analysis, ANN.*

1. INTRODUCTION

Human beings suffer from a wide variety of chest-related diseases. These chest diseases include asthma, copd, pneumonia, tuberculosis, etc. The chest diseases have symptoms that demonstrate their presence. Symptoms include shortness of breath, chest congestion, chest pain, cough from the throat, and cough from the chest, etc. and manifest in difference, these are the common symptoms which are found in many situations. When human beings do regular functions in their day to day lives, they are prone to seek these symptoms in situations such as running, walking, long breathing up, etc. To detect which chest disease the human being might be facing, a plan is identified by which decision can be made by making use of a symptom-based questionnaire. To make the machine understand and predict which disease the patient suffers from, it must be trained on the sample datasets containing symptoms in questionnaire. Such datasets can be obtained from UCI database, CHHS (California health and human services) database, as well as data from reputed national institute of tuberculosis and respiratory diseases.

A large number of people who suffer from chest related diseases die due to wrong predication of chest conditions. This is often due to the fact that they are diagnosed much later after the disease occurs, after which it becomes difficult to solve the problem. In addition to this, they are often misdiagnosed for one another. A patient with Asthma may be told he has COPD and vice versa since there is a very thin line difference between these two diseases. Initially they are so identical that hardly difference is there. This leads to the wrong treatment being given to the patient and causes adverse effects of the treatment. Therefore, there is a need to build an easy system to aid doctors for preliminary decision making. There is also a need to empower the patient with a tool that helps him understand his condition better and take appropriate measures by giving proper information of his condition of health to the correct doctor.

Mainly focus is on collection of information for Knowledge Discovery in Databases (KDD). This is initial process from which mashup candidates are identified by addressing a repository of open services. Within this approach, there is a customized approach to life cycle which software engineers can use to generate new applications based on service integration techniques. KDD also define service integration qualification by discovering different aspects of web service specifications.

Volume 1, Issue 2 (2019)
Article No. 14
PP 1-4

2. AVAILABLE METHODS

a. Support Vector Machine (SVM)

Support vector machine is a supervised learning model that is defined as the finite dimensional vector spaces where each dimension characterizes a feature of a particular object. In this way, SVM has been proved as an effective method in high-dimensional space problems. Due to its computational competence on huge datasets SVM is typically used in document classification, sentiment analysis and prediction-based tasks

b. K-Nearest Neighbors (KNN)

K-Nearest Neighbor (KNN), a supervised learning model as well, is used to classify the test data using the training samples directly. In KNN, an object is classified by the majority voting of its closest neighbors. Alternatively, the class of a new sample is predicted based on some distance metrics where the distance metric can be a simple Euclidean distance. In the working steps, KNN first calculates k (No. of the nearest neighbors). After that, it finds the distance between the training data and then sorts the distance. Subsequently, a class label will be assigned to the test data based on the majority voting.

c. Artificial Neural Network (ANN)

The Artificial Neural Network (ANN), also a supervised learning strategy, contains three layers: input, hidden and output. The connection between the input units and the hidden and the output units are based on relevance of the assigned weight of that specific input unit. Usually, if the weight is higher, then it is considered more important. ANN may use linear and sigmoid transfer (activation) functions. Also, the ANNs are suitable for the training of large amounts of data with limited inputs. For multi-layer feed forward ANN, the mostly used learning algorithm is the Backpropagation learning tool. In ANN, the input data records should be separated into three sub-datasets for the purpose of training, validation and testing.

3. PROPOSED METHOD

Symptom-based Questionnaires required: Asthma, COPD, Pneumonia, Tuberculosis.
Training of machine using datasets:

a. Dataset required for the purpose can be obtained from UCI repository database, CSSH database
b. Datasets from the National Institute of Tuberculosis and Respiratory Diseases (India).
c. An ML training service like Tensor Flow can be used to train the system based on the dataset selected.
d. A Cloud ML service can be used to verify and double check the training.
e. Predictive analysis is carried out to find the % of accuracy diagnosis for a particular disease

The main steps which are considered when a predicated disease is to be notified.
Step 1. Collection of user data
Step 2. User choice
Step 3. Collection of questionnaire
Step 4. Processing of data collection
Step 5. Comparing of the data received with data set
Step 6. Result of comparison decision to be taken with respected to which chest disease
Step 7. Depending on the decision proceed for treatment, if ok update the data set

Currently systems utilize a large amount of medical data taken from tests that determine the nature of the chest disease. These are expensive and not scalable in nature and require advanced medical professionals. To overcome problems on existing system, in proposed system user does not require to search data in various repository with special features. User need only to give information which is required to be collected. User can just type combination of queries and based on user behaviour analysis exact data will be predicted.

However, 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 symptoms based questionnaires for the broader public. We seek to integrate several of these symptom-based questionnaires with real life scenario data to e 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 normal people with no chest conditions. 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.

Volume 1, Issue 2 (2019)
Article No. 14
PP 1-4

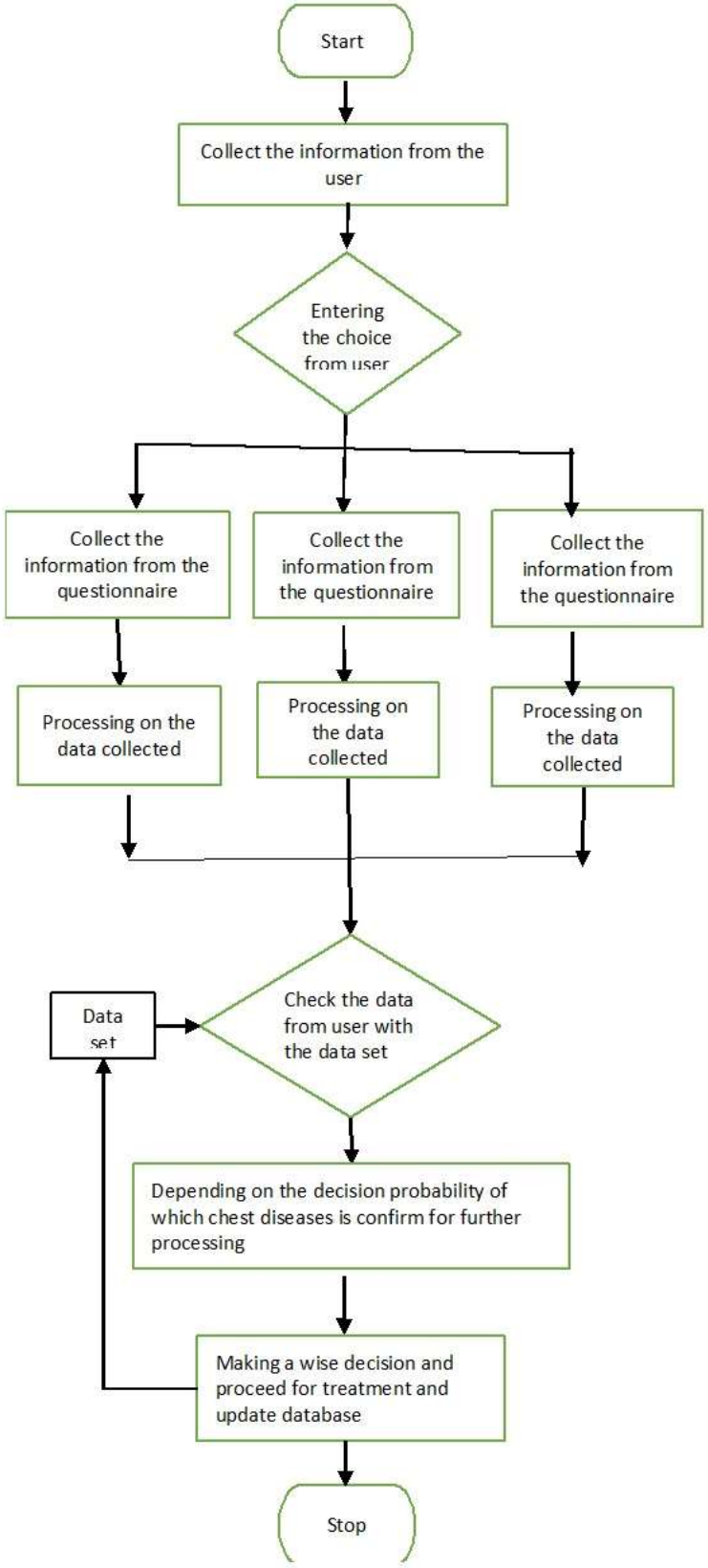

Fig 1. Flow chart.

The main focus of Predictive Diagnosis System will be to implement machine learning algorithms and Prediction of which chest disease the patient might be suffering from based on the symptoms.

Volume 1, Issue 2 (2019)
Article No. 14
PP 1-4

4. ANALYSIS AND ADVANTAGES

There are important applications for this type of systems there are:

a. In hospitals for doctors to use as an initial diagnosis measure before further check-ups.
b. For self-diagnosis by patients
c. By government and municipal bodies to see the impact of air pollution on health of citizens.
d. Industries to ensure health and welfare of people before setting up manufacturing and other plants near occupied localities.

Major advantage of the proposed system is that it is generally very difficult to predict chest diseases other than heart disease as data and specific criteria's to diagnose chest diseases are not available. All the system which are available currently are focusing only on the heart disease prediction. Periodic record of PFT (Pulmonary function test) gives regular input about the patient's condition. This system can predict COPD and Asthma well in their initial stages. COPD and Asthma can be controlled if diagnosed in initial stage itself. Whereas Pneumonia and heart disease can even be diagnosed and treated as substantial research has already been done.

5. CONCLUSION

A prototype chest disease prediction system is developed using three data mining classification Modeling techniques. The system extracts hidden knowledge from a historical chest disease database. DMX query language and functions are used to build and access the models. The models are trained and validated against a test dataset. Lift Chart and Classification Matrix methods are used to evaluate the effectiveness of the models. All three models are able to extract patterns in response to the predictable state. The most effective model to predict patients with chest disease appears to be Artificial Neural Network and Decision Trees. The goals are evaluated against the trained models. All three models could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. This system can be further enhanced and expanded. It can also incorporate other data mining techniques, e.g., Time Series, and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases.

REFERENCES

[1] 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. ***"A New Symptom Based Questionnaire for predicting the presence of Asthma"*** 2006;73:296-305. doi: 10.1159/000090141

[2] 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 2006;73:285-"***Symptom-Based Questionnaire for Differentiating COPD and Asthma",*** 295. doi: 10.1159/000090142

[3] 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 ***"Symptom-Based Questionnaire for Identifying COPD in Smokers, Respiration"***

[4] S. Fathima and N. Hundewale, "**Comparison of Classification Techniques- Support Vector Machines and Naive Bayes to predict the Arboviral Disease-Dengue**," IEEE International Conference on Bioinformatics and Biomedicine Workshops, 2011.

[5] S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan and T. Zhu,Cardiovascular "**Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework**," International Conference on Big Data Analysis, 2017

[6] S. Pouriyeh, S. Vahid, G. Sannino, G. D. Pietro and H. Arabnia, J. Gutierrez, "***A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease,"*** IEEE Symposium on Computers and Communication, 2017.

[7] 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) ***"A Tree-based Decision Model to Support Prediction of the Severity of Asthma Exacerbations in Children"***

[8] Vinayak Singh, Anant Gaikwad, Surendra Waso, Eknath Sawale, IJIRCCE Vol 4, Issue 3, March 2016, PP3253-3258 ***"Web Based e-Health Systems and Services,"***

[9] Statlog database: http://archive.ics.uci.edu/ml/machine-learningdatabases/statlog/heart/

[10] Cleveland database: http://archive.ics.uci.edu/ml/datasets/Heart+Disease