DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

A MACHINE LEARNING BASED APPROACH TOWARDS DISEASE PREDICTION AND MEDICINE RECOMMENDATION FOR ANY VIRAL DISEASE.

SUBMITTED BY:

T. HARSHINI AP18110010264

S. MADHU PRIYANKA AP18110010265

M. SUPRAJA RAMA AP18110010301

P. DIVYA KALYANI AP18110010325

UNDER THE ESTEEMED GUIDANCE OF:

Dr. RAVI KANT KUMAR

(Assistant Professor)

ABSTRACT

Viruses and humans have been at odds for millennia. Viruses that transmit infection are rapidly emerging day by day. People nowadays suffer from a variety of diseases as a result of the climate and their lifestyle choices. Many of the determinants of global health have faltered, despite the fact that many generally averaged indicators of population health have increased. Many new infectious diseases have arisen, although some have reappeared unexpectedly. COVID-19, a novel coronavirus discovered recently, is an example. No medication was available to treat the patients. As a result, predicting illness at an earlier level becomes a crucial challenge. However, a doctor's ability to make an appropriate diagnosis based on symptoms is hampered. A large number of infected people have been exposed to the virus during this process. Owing to a lack of enough medical treatment, some died. As a result, we've incorporated a system in this project. We developed a general disease prediction system focused on the patient's symptoms. Furthermore, this approach examines the symptoms and suggests the appropriate medicine to cure the predicted viral disease by using Machine Learning techniques. In brief, this paper presents an application programming interface for recommending medications to users that are suffering from a certain illness, which the system can also diagnose by evaluating the patient's symptoms using machine learning algorithms.

Keywords: Machine Learning techniques, viruses, symptoms, prediction, recommendation

1. INTRODUCTION

Every year, terabytes of data are produced in the healthcare industry. Health records serve as a repository for patient data. Extracting reliable knowledge or providing high-quality treatment is a difficult and crucial job [1]. Every two months, it is estimated that more than 70% of Indians are susceptible to viral fever, cough, cold, and other common body maladies. Since, many people are unaware that general body ailments could be side effects to anything more harmful, 25% of the population succumbs to death as a result of missing early general body symptoms [2]. As a result, an intelligent device for disease prediction plays a critical role in disease prevention and sustaining people's health by ensuring precise and reliable disease risk prediction [3].

Standard body diagnostics are now needed in order to stay healthy. There are numerous sources available today as individual prediction or advisory systems, but the need of the hour is for an interconnected model that includes both. This interface is a user support mechanism that allows patients to get information about their illness as well as medications that are relevant to that disease. In the proposed work, the focus is on machine learning techniques to extract hidden rules and relationships between symptoms and diseases. Patient will enter the various symptoms in the user interface of the application where the data will be processed and the most probable disease is predicted based on the symptoms. Our work also recommends the medicine based on the predicted disease.

Further part of this document is organized as: Section 2 describes the related works. Section 3 discusses about proposed methodology. Implemented techniques, results and obtained accuracy are described in Section 4. Section 5 shows system overview. GUI is shown in Section 6. Finally, the work is concluded in Section 7 and future scope is also included.

2.RELATED WORK

Research has been done by many people to build models which can predict diseases. In this section the previous related studies are reviewed.

Thakkar, B.A. et al [4], Healthcare decision support system for swine flu prediction using naïve bayes classifier, focuses on the aspect of medical diagnosis by learning patterns through the collected data for swine flu using naïve bayes classifier for classifying the patients of swine flu into three categories (least possible, probable or most probable), resulting into an accuracy of nearly 63.33%. Datasets used for this classification were limited in number.

Jackins, V. et al [5], Health care decision support system for swine flu prediction using naïve bayes classifier. Discovered a model for the diagnosis of diabetes, coronary heart disease and cancer among the available dataset. The dataset is chosen from online repositories. The techniques of pre-processing applied are filled in missing values and removing correlated columns. Next, the classifier is applied to the pre-processed dataset, and then Bayesian and random forest models are constructed. Finally, the accuracy of the models is calculated and analyses are based on the efficiency calculations.

Mohan, S. et al [6], Effective heart disease prediction using hybrid machine learning techniques. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease.

Gupta J.P. et al [5], A Computer-based disease prediction and medicine recommendation system using machine learning approach, developed disease prediction and medicine recommendation using various machine learning algorithms like Naïve Bayes, Decision Tree and Random Forest. The system has been trained by mapping the various symptoms of the diseases in the dataset. Disease prediction level (High, Average and Low) has also been analysed in this work.

3.PROPOSED WORK

A system for disease prediction and medicine recommendation plays a crucial role in effective treatment. The main objective of this work is to develop an interface which uses some machine learning algorithms, languages used will be python. The system takes the symptoms from the users which they are feeling at that moment, runs these Machine Learning algorithms and predicts the most accurate disease followed by medicine recommendation.

3.1 DATA GATHERING

The raw data is gathered from websites like mayoclinic.org, healthgrades.com, medicines.org.uk, healthline.com, dataworld.org, cdc.gov and kaggle.com. It is a CSV file containing four columns, namely diseases, related symptoms for that particular disease, drug and its composition. All the symptoms for the particular disease are recorded and the data is then sent for the pre-processing so that the python code can be implemented efficiently.

S. No.	DISEASE	SYMPTOMS	RECOMMENDED MEDICINE
1.	Cytomegalovirus	Fatigue, Swollen glands, Fever, Sore throat, Muscle aches.	Cidofovir, Foscavir, Ganciclovir, Valganciclovir
2.	Yellow Fever	Fever, Abdominal pain, Vomiting, Sluggishness, Jaundice, Lethargy	Stamaril
3.	HPIVs	Runny nose, Ear pain, Sore throat, Barky Cough, Wheezing, Fever	Ribavirin, Acetaminophen, Ibuprofen
4.	Viral Meningitis	Fever, Headache, Sleepiness, Nausea, Vomiting, Loss of appetite	Acyclovir
5.	Infectious Mononucleosis	Fatigue, Fever, Sore throat, Body aches, Enlarged lymph nodes	Acyclovir, Valacyclovir

Table 1: Collected Data

The dataset contains 5088 rows contains multiple diseases that counts up to 41 unique disease and for the drug dataset each disease has multiple drugs. All the diseases present in the dataset are included in the Table below.

Cytomegalovirus	Yellow Fever	HPIVs	Viral Meningitis
Infectious Mononucleosis	Shingles	Chicken Pox	Swine Flu
Pneumonia	Measles	Hepatitis B	Chikungunya Virus
Cholera	Tuberculosis	Human Papilloma Virus (HPV)	Herpes
Bird Flu	Small Pox	Gastroenteritis	Monkeypox Virus
Zika Virus	Scarlet Fever	Dengue	Diphtheria
Cryptosporidiosis	Whooping Cough	Fifth Disease	Molluscum Contagiosum
Rhino Virus	Encephalitis	Typhoid	Anaplasmosis
HIV	Rosea	Lassa Virus	Common Cold
Warts	Viral Myocarditis	SARS	Influenza
HIB disease	Hepatitis C		

Table 2: List of all diseases

After gathering the data in raw form, the data is transformed into another .csv file which has indexing of every symptom mapped with every disease in 0 and 1's by the method of one-hot encoding. The diseases are made as rows and all the symptoms are made as columns like the table made below, the presence of every symptom for a particular disease is marked as 1 and its absence is marked as 0 all according to the dataset which we collected earlier i.e., raw dataset. From table [3] we can infer that the symptoms like fever, Fatigue are present in Cytomegalovirus that's why it is marked 1 and since Cough and Vomiting are not present so it is marked as 0 according to the dataset which we collected and same goes with Yellow Fever, Swine Flu and Cholera.

Disease	Fever	Fatigue	Cough	Vomiting
Cytomegalovirus	1	1	0	0
Yellow Fever	1	0	0	1
Swine Flu	1	0	1	1
Cholera	0	1	0	1

Table 3: Cleaned Data

4. IMPLEMENTATION TECHNIQUES

We used three different algorithms for the diagnosis of the disease like the Decision tree, Random forest and Naïve Bayes. For these classification models we created all combinations of disease and symptom as we have one to one mapping of the disease and symptom. We have trained our disease prediction system using all the three classifiers separately.

4.1 Decision Tree

The Decision Tree methodology is a widely used machine learning technique for defining classification and prediction structures based on several explanatory parameters in order to create prediction models for a target instance. This path classifies a population into branch-like segments in a tree that construct an inverted tree with a root node, internal nodes, and leaf nodes [8].

The Decision Tree predicts a disease based on the underlying symptoms. To begin, we collect the user's top five symptoms and store them in an array with the value 1 allocated to each value. This is fed into the disease prediction model as an input. This array corresponds to the disease data set and finishes with the highest level of confidence at a typical leaf node. And finally, the disease is predicted.

sklearn kit is used to implement a decision tree algorithm in python to predict the disease according to the symptoms given. We created the test, train data from the collected dataset and used accordingly. The data is been read using variable_name=read_csv()and training is done by implementing the DecisionTreeClassifier() function and to fit the model using model.fit().

4.2 Random Forest

Random forest algorithm is one of the most effective classification approach. It is a hierarchical collection of tree structured base classifiers. RF algorithm has been used in probability estimation and prediction. RF consists of many decision trees [9].

First the algorithm chooses random data samples from dataset and construct decision trees for every sample data set chosen. At every step the predicted result is voted and most voted prediction is taken as result of this classification.

4.3 Naïve Bayes

A naive Bayes classifier is a concept that refers to a basic probabilistic classification based on Bayes' theorem. A naive Bayes classifier implies that the presence (or absence) of one feature of a class is irrelevant to the presence (or absence) of any other feature [10].

We used python language for the implementation of it. sklearn kit is imported and loaded. Then the dataset is been read using read_csv(). All of this done is using NaiveBayes() and predict() present in sklearn package. The accuracy is more when the whole dataset is trained.

5. SYSTEM OVERVIEW

A basic user interface is built for the system which acts as a supporting tool for the patients to predict the disease and prescribe medicines. Through this interface we can give five symptoms which are used to predict the disease and shows the medicine that can be used to treat the disease. The system working overview is shown in Fig.1.

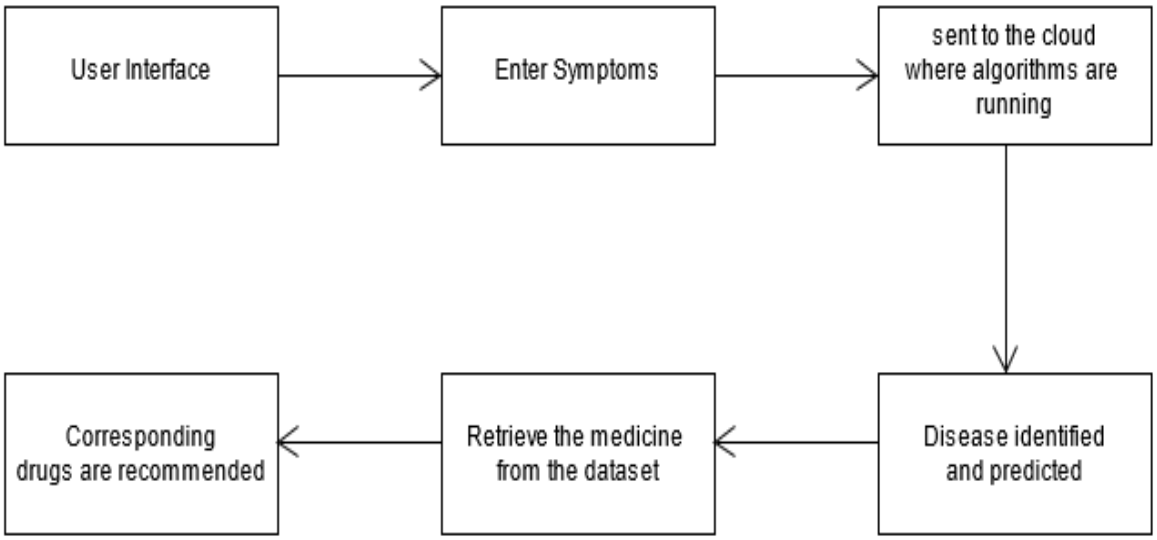

Fig.1. Proposed System

6.RESULTS AND DISCUSSION

We have developed GUI where users can enter the symptoms and can predict the disease. It also recommends the medicine to cure that particular disease. After building the model we got the classifiers scores as shown in Table 4. GUI is shown in fig 2 and fig 3.

MODEL	CLASSIFIER SCORE
DECISION TREE	0.92
RANDOM FOREST	0.99
NAÏVE BAYES	0.99

Table 4: Classifier Scores

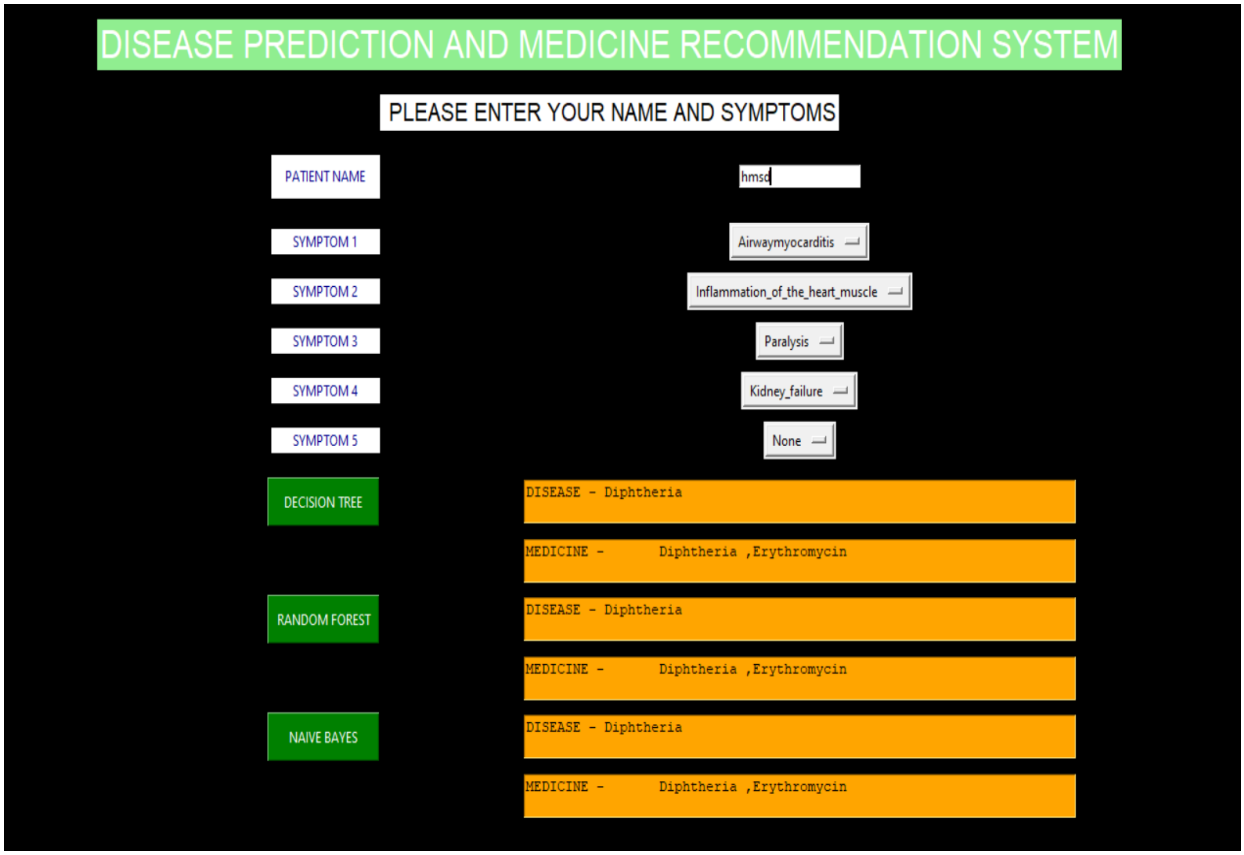

Fig2: GUI of disease prediction and medicine recommendation system

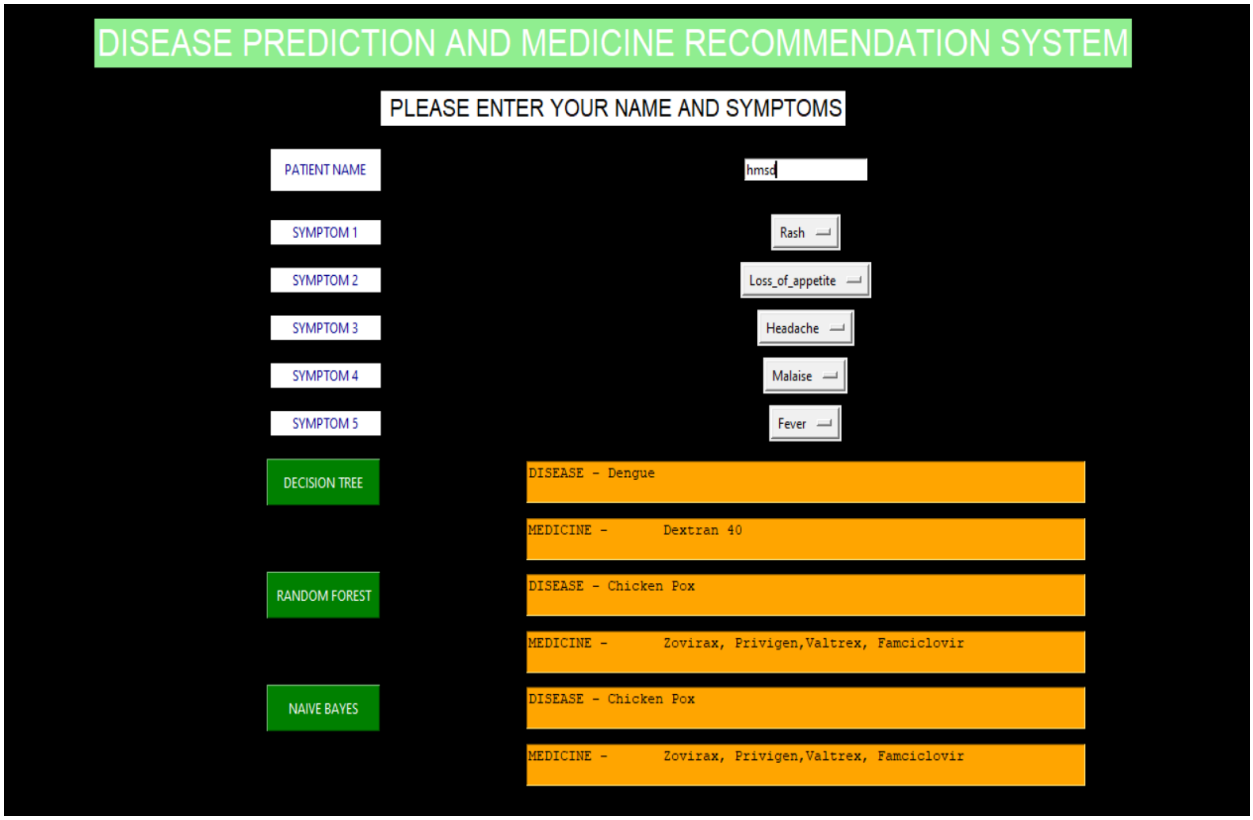

Fig3: GUI of disease prediction and medicine recommendation system

7.CONCLUSION AND FUTURE SCOPE

We set to develop a system that predicts the disease based on symptoms given to it and recommends the appropriate medicine. We were successful in developing such a system using various machine learning algorithms like Decision Tree, Random Forest and Naive Bayes. In this process we have also developed GUI so that it can be very easy to be used by the patients. It has visual representation of data collected and results achieved.

The features of this system are many like paper free work, automation of disease diagnosis, maintaining the information related to healthcare etc. So, this system has large scope. We work further to obtain accurate results. And the system can be improved further by including many other symptoms, training and testing them. And also, after classifying and analysing the work we have done, we would like to help drug experts in development of new and effective medicines from the observations we have done earlier.

8.REFERENCES

[1] Pandey, S.C., 2016, October. Data mining techniques for medical data: a review. In 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES) (pp. 972-982). IEEE.

[2] Singh, M., Martins, L.M., Joanis, P. and Mago, V.K., 2016, July. Building a cardiovascular disease predictive model using structural equation model & fuzzy cognitive map. In 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1377-1382). IEEE.

[3] Gujar, D., Biyani, R., Bramhane, T., Bhosale, S. and Vaidya, T.P., 2018. Disease prediction and doctor recommendation system. International Research Journal of Engineering and Technology (IRJET), 5, pp.3207-3209.

[4] Thakkar, B.A., Hasan, M.I. and Desai, M.A., 2010, October. Health care decision support system for swine flu prediction using naïve bayes classifier. In 2010 International Conference on Advances in Recent Technologies in Communication and Computing (pp. 101-105). IEEE.

[5] Jackins, V., Vimal, S., Kaliappan, M. and Lee, M.Y., 2021. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing, 77(5), pp.5198-5219.

[6] Mohan, S., Thirumalai, C. and Srivastava, G., 2019. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, pp.81542-81554.

[7] Gupta, J.P., Singh, A. and Kumar R.K., 2021, March. A Computer-based disease prediction and machine recommendation system using machine learning approach. International Journal of Advanced Research in Engineering and Technology (IJARET), 12, pp.673-683.

[8] Radhika, S., Shree, S.R., Divyadharsini, V.R. and Ranjitha, A., 2020. Symptoms Based Disease Prediction Using Decision Tree and Electronic Health Record Analysis. European Journal of Molecular & Clinical Medicine, 7(4), pp.2060-2066.

[9] Jabbar, M.A., Deekshatulu, B.L. and Chandra, P., 2016. Intelligent heart disease prediction system using random forest and evolutionary approach. Journal of Network and Innovative Computing, 4(2016), pp.175-184.

[10] Pattekari, S.A. and Parveen, A., 2012. Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences, 3(3), pp.290-294.