Analysis of Disease Prediction for Common Behavioral Symptoms

Angel Jain[1], **Aditya S Chandel**[2] and **Dr. Ajay Kumar**[3]
[1]Optum, Bangalore, India, [2]TCS, Noida, India, [3]iNurture Education Solutions, Delhi, India
[1]jainangel1999@gmail.com. [2]aditya.pranshu00@gmail.com,
[3]kumarajay7th@gmail.com

Abstract -- **Late diagnosing of diseases is the main reason for a huge number of deaths in the world over the last few decades. There is dire need for a reliable, accurate, and feasible system to diagnose diseases in time for proper treatment. To automate the analysis of large and complex data, machine learning algorithms and techniques have been applied to analyze various medical datasets. Recently, researchers used machine learning techniques for diagnosis of diseases. This paper presents a survey of various models based on such algorithms and techniques and analyzes their performance. Models based on supervised learning algorithms such as K-Nearest Neighbour, Naïve Bayes, Decision Trees and Random Forest are found popular.**

Keywords: Dignosing of Diseases, Machine Learning, K-Nearest Neighbour, Naive Bayes, Decision Trees, Random Forest

I. INTRODUCTION

GOOD health is an important part of our life that helps us to keep active. If our health is not good, then the concentration of our body is affected and sometimes we get depressed, and sometimes a disease is not diagnosed correctly which in a severe case can lead to demise. Change in lifestyle, work-related stress, and consumption of unhealthy food or substance may lead to bad health for a person.

Globally, medical organizations collect data on various health-related parameters. These data can be utilized via machine-learning techniques to gain useful insights. Nevertheless, the data collected is enormous and, many times, it can be quite noisy. These datasets, which are too overwhelming for human minds to comprehend, can be conveniently explored using various machine-learning techniques. Accordingly, such algorithms have become very useful to predict the presence or absence of health-related diseases accurately.

II. LITERATURE SURVEY

Kourou *et al.* [1] proved that the integration of multidimensional heterogeneous data combined with the application of different techniques for feature selection and classification provides promising tools for inference in cancer prediction research. They proposed a new method for the detection of skin cancer using an ANN. More clinical decision support systems can be developed by introducing the analysis of genes in the prevalent ML algorithms. Algorithms like Linear Regression, Support Vector Method (SVM), Kernel SVM, Naïve Bayes, Decision Trees, Random Forest Classification, and K-Nearest Neighbors can be used for accurate decision prediction. Using the K-fold cross-validation technique will provide more accuracy and also help in determining the important parameters for the algorithms.

Vijayalakshmi and Kumar [2] presented the data mining concept of "Disease Prediction by using Machine Learning". Anjali Bhatt *et al.*[3] reported the technique of using ML to predict diseases. The concept of machine learning is applied to disease-related information retrievals and the treatment processes are achieved by using data analysis. The predictions of outbreaks in diseases are using the decision tree as it is quite effective. This concept predicts the results at low cost and low time.

Varma and Senthil [4] extract a feature vector for each new document by using feature weighting and feature selection algorithms for enhancing text classification accuracy. Thereafter, classifier is trained by Naïve Bayesian (NB) and support vector machine (KNN) algorithms. In experiments, both algorithms showed acceptable good results for text classification.

Pareek *et al.* [5] presented the concept namely, "Disease prediction using Machine Learning over Big Data". Big data has following features:

(i) medical data analysis with accuracy,
(ii) early prediction for disease,
(iii) patient-oriented data with accuracy,
(iv) The medical data, is securely stored and used in many places.
(v) incomplete regional data are reduced, giving more accurate results.

III. PROPOSED SYSTEM

Machine learning algorithms namely Naive Bayes, Decision Tree and Random Forest Classification algorithm were used to

This paper was presented during the Poster Session of the International Conference on 'Advances and Key Challenges in Green Energy and Computing', organised by Ajay Kumar Garg Engineering College, Ghaziabad during 24-25 February 2023.

classify and predict the disease according to the symptoms and will also predict the correct medicine for that disease shown in Figure 1. In the very 1st step of the proposed model, the user data or patient data are used to put the data to the system. 2nd step is to select or scan the symptoms of the data received from the patient. 3rd step is to answer some queries based on the patient conversation queries. 4th step tells machine learning model to predict the best outcome as disease diagnosed. 5th step analyses the result received from different machine learning algorithms and best probability is filtered out based on their historical questions.

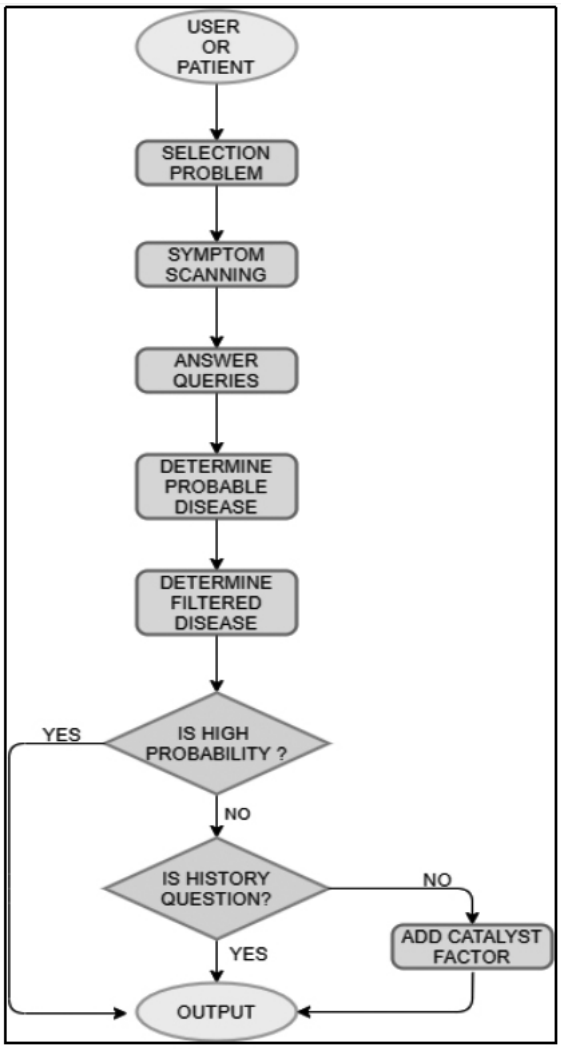

Figure 1. Proposed Flowchart.

IV. IMPLEMENTATION

In this section, an experiment has been performed with the two datasets using various machine learning algorithms. The first dataset [6] contains 41 rows and 156 columns. The second dataset [7] contains 4921 rows and 156 columns.

Data File: Disease Prediction and Medication Dataset is a hybrid dataset. It contains 2 datasets. Both datasets contain the symptoms of many diseases and their medicines. The symptoms for the specific 100 diseases have been collected and is being used by Google and many other health websites. The main symptoms are Allergy, Chronic Cholestasis, Drug Reaction, Fungal, GERD and Peptic Ulcer Disease.

The common diseases in humans are: Diabetes, Jaundice, Hepatitis C, Hepatitis D, Piles, Arthritis, Gastroenteritis, Malaria, Heart Attack, Paroxysmal Positional Vertigo, Bronchial Asthma, Chickenpox, Hepatitis E, Varicose Veins, Acne, Hypertension, Dengue, Alcoholic Hepatitis, Hypothyroidism, Urinary tract infection, Migraine, Typhoid, Tuberculosis, Psoriasis, Cervical Spondylosis, Hepatitis A, Common Cold, Hypoglycemia, Impetigo

Decision Tree: It is a classification technique used to predict categories for new events and it helps to classify data. The necessary approaches to building a decision tree are-ID3, C4.5, and CART.

Step 1: The expected information (Entropy): we need to classify: a pattern in dataset D which can

be calculated by:

$$\text{Entropy, Info(D)} = \sum_{i=1}^{m} Pi \log_2(Pi) \quad (1)$$

Step 2: Then we calculate Information Gain as: -

$$\text{Gain(A)} = \text{Info(D)} - \text{lnfoA(D)} \quad (2)$$

The maximum accuracy obtained from the decision tree is 96%.

Random Forest Classification: In this learning, we apply the algorithm multiple times to increase its power. It is a version of ensemble learning. In this classification, we use some part of decision tree classification. The maximum accuracy obtained is 95%.

$$MSE = 1/N \quad (3)$$

$$\text{MSE} = \left(\frac{1}{N}\right) \sum_{i=1}^{N} (\text{f}_\text{i} - \text{y}_\text{i})^2 \quad (4)$$

where N is the number of data points,
f_i is the value returned by the model and
y_i is the actual value for data point i.

Naïve Bayes This algorithm requires some independent assumptions. It comes under supervised learning. It divides the data points into different categories and then it finds the probability of the different categories. The category with the highest probability is considered the most likely one for the

new data point. The probability is calculated as follows

$$P(A/B) = \{P(B/A) * P(A)\}/P(B) \quad (5)$$

The maximum accuracy obtained is 99%.

V. RESULTS

Machine learning algorithms were investigated in this paper to predict diseases. Accuracy achieved via using the Naïve Bayes algorithm is 99% while that obtained using Decision tree algorithm is 96% and 95% from the Random Forest algorithm. The result are depicted in Figure 2.

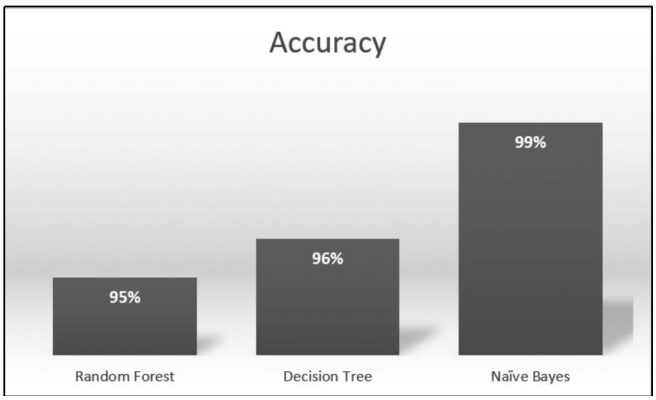

Figure 2. Comparisaon of Accuracy results.

VI. CONCLUSION

This paper investigated a technique to predict a disease based on its symptoms. The experiment has been performed in such a way that the system takes symptoms from the user as input and produces output. *i.e.* predicts disease. Disease predictor was successfully implemented using grails framework. Machine Learning based algorithms are applied and Naïve Bayes is found to be the most effective algorithm.

REFERENCES

[1] K. Kourou, T. P. Exarchos, K.P.Exarchos, M.V. Karamouzis and D.I. Fotiadis, "Machine learning applications in cancer prognosis and prediction", *Computational and Structural Biotechnology J.*, vol. 13, 2015, pp.8–17, Elsevier.

[2] C.S. Vijayalakshmi and Niharika Kumar, "Survey on risk estimation of chronic disease using machine learning", *Int'l Research J. Engineering and Technology*, vol.6, no.6, 2019.

[3] Anjali Bhatt, Shruti Singasane and Neha Chaube, "Disease prediction using machine learning", *Int'l Research J. Modernization in Engineering Technology and Science*, vol.4, no.1, 2022, pp. 310-314.

[4] Bhavitha Varma and B. Senthil, "A different type of feature selection methods for text categorization on imbalanced data", *J. Network Communications and Emerging Technologies*, vol. 8, 2017.

[5] Deepika Pareek, Manish Tiwari, Mayank Patel and Amit Sinhal, "Heart disease prediction using data mining", *Int'l J. Innovative Research in Science, Engineering and Technology*, vol. 8, no.8, 2019, pp.8469-8479.

[6] https://github.com/anujdutt9/Disease-Prediction-from-Symptoms/tree/master/dataset

[7] https://github.com/yaswanthpalaghat/Disease-prediction-using-Machine-Learning.

Dr. Ajay Kumar is currently working as a Mentor at iNurture Education Solution, Delhi, India. He received his PhD degree in CSE from DIT University Dehradun, India in 2021. He received B.E. (ISE) from SJCE Mysore (now JSSUTA Mysuru) Karnataka in 2005 and subsequently M.E. (SE) from Birla Institute of Technology (BIT), Mesra, India in 2007.

He is an inspiring and energetic technocrat with 15 years of research in academia and industry experience. His research interests include Machine Learning, Deep Learning, NLP, Computer Vision etc. He has published over 22 papers in reputed international/national journals and conferences. He has organized and attended various workshops during his academic career. He is a member of IEEE, ACM and CSI.

Aditya Singh Chandel is currently employed at TCS Noida. Over the past two years, he gained valuable experience and honed skills in various NLP techniques and machine learning. He holds a B.Tech. (CSE) degree from DIT University Dehradun, India. Throughout his academic journey, he developed a keen interest in exploring the vast possibilities of machine learning and its applications. He got actively engaged in numerous machine learning projects, as well as Python UI-based projects. He is passionate about leveraging knowledge and skills to contribute to the ever-evolving field of technology. With a firm grasp of NLP techniques, ML methodologies, and a strong foundation in computer science, he is ready to take on new challenges and make a meaningful impact in the world of software engineering.

Angel Jain, is currently working at Optum Global Solution, Bangalore as Associate Software Engineer II. He obtained B.Tech. degree in CSE from DIT University, Dehradun, India in 2021. Throughout his academic journey, he cultivated a strong passion for delving into the expansive realm of machine learning and its diverse applications. In 2021, he had the opportunity to work as a Programmer Analyst Trainee Intern at Cognizant from March-July'21. Thereafter, he joined Optum as an Associate Software Engineer II in August 2021. With nearly two years of experience in the field of Data Warehousing, he honed his skills as an ETL developer and gained proficiency in Reporting and scripting languages. He aims to leverage his enhanced capabilities to build a fulfilling career and contribute value to the industry he is associated with.