



International Journal of Marketing Management

ISSN 2454 - 5007



www.ijmm.net

Email ID: editor@ijmm.net , ijmm.editor9@gmail.com

Human Symptom-Based Disease Prognosis

, ¹Syed Idrees, ²MOHD OMER FAROOQ, ³SHAIK FAISAL, ⁴RAFI AHMED,

¹Assistant Professor, Department of CSE-AIML, Lords Institute of Engineering & Technology.

^{2,3,4} Student, Department of CSE-AIML, Lords Institute of Engineering & Technology.

Abstract

Modern medicine has led to the discovery of new illnesses with similar or even more severe symptoms than those of earlier eras; nonetheless, these diseases pose a considerably greater threat to human health. So, this approach can detect new illnesses by listening to the user's input, then recognizing the condition and showing it on the user's smart screen. One of the main goals of this effort is to find a way to prevent diseases before they kill people. The illness prediction algorithms used include Naive Bayes and Random Forest. In order to make this forecast, we compare and contrast the two algorithms' accuracy levels and then use the one with the higher level of accuracy to forecast the illness. Python illnesses finish its implementation. Using just machine learning and the Python programming language, this goal is symptom-based. Using the Python T-kinter, the user interface was created. The computer may be taught to forecast the illness utilizing the prior data set that is accessible in hospitals. computer language and the tkinter package.

Keywords—Random Forest, Diagnosis of disease, Prediction, Machine Learning Algorithm, Database

INTRODUCTION

In this chapter, we will go over the background of the issue, how it was described, some definitional basics, and several applications of object detection. 1.1 Where the Issue Started In modern medicine, the function of the health care department in treating patients' illnesses is crucial. Therefore, the healthcare sector must provide the highest level of safety in addition to the advantages for the individual. Many individuals aren't paying enough attention to their health because of the world's time and money constraints; in fact, 40% of people forget about illnesses that might develop into serious problems down the road. Modern scientific medical

practitioners use a wide variety of clinical tools and methodologies to detect and diagnose illness. Consistently good prognosis is the key to effective therapy. If the patient is unable to see a doctor or health center, they may still learn what ailment they're dealing with by inputting their symptoms and other relevant information. Health organizations may benefit more from this study if patients just fill out a short form on a website or mobile app and report their symptoms. When the time comes, the gadget can provide precise o o This initiative was developed in response to data showing that 25% of Indians afflicted with common illnesses are losing their lives due to delayed diagnosis, and 70% of those persons suffer from common illnesses. Users may access the interface from any location, allowing them to conduct exams from the comfort of their own homes or offices. GUI is built in this type of method so that everyone can readily perform by utilizing it and may check their health. 1.2 Overview and Foundational Terms Random Forest (1.2.1): As a supervised M.L. method, random forest finds widespread use in the domains of regression and classification. It uses a sampling approach to choose samples at random, builds decision trees for each sample, and then takes the decision trees' majority votes into account. Because it can process facts with continuous values, numerical values, and categorical values for classification, the RF Algorithm's rule set is one of its most important characteristics. When it comes to categorization models, it works better. The Random Forest algorithm's operation: o Step 1: Picking random samples from a dataset is the first step. Step 2: After that, for each randomly chosen pattern, this set of rules will build a decision tree. Afterwards, it will get the anticipated outcome from every selection tree. o Step 3: At this stage, voting may be completed for all expected outcomes. Fourth, before you close, choose the most popular forecast and announce the final prediction outcome. section 1.2.2 The term "naïve bayes" is used to describe a classification system.

information obtained mostly from the Bayes Theorem, which is based on the probability of the

specified classifications. Algorithms that utilize probability to find outcomes in the absence of specified class labels in the training dataset are used for data without class labels. Neural Bayes Classifier Operation:

- o First, make frequency tables from the provided dataset.
- o Step 2: Build a probability table by identifying the possibilities of the provided skills.
- o Thirdly, determine the posterior probability statement by using the Bayes theorem.

Issue Statement with Outcomes and Objectives (1.3) 1.3.1 Issue Description: There are some issues with machines or devices in the health industry that produce unacceptable results. To avoid these problems and achieve the desired results, we are developing a mission to provide accurate predictions using the data provided by the user and the datasets available in the system. There may be a lot of work to be done in this business, which might be rather large. In order to help those in need, this task was completed using various methods, tactics, and approaches. The project's creators set intended to make it easier for those in need to understand their health status; for example, if a character has been experiencing certain symptoms but isn't sure what illness they're dealing with, the project may provide some guidance. This might potentially cause a range of diseases down the road. Predicting the signs and symptoms of this illness will be very helpful for many people, from children and teenagers to adults and the elderly, so they can avoid becoming sick and learn about it early on. 1.3.2 Goals: 1. The main goal of this task, called "Disease Prediction based symptoms," is to identify the correct illness a person is suffering from by analyzing all of their symptoms and indicators. 2. A graphical user interface is made available for the user to input their symptoms by means of this challenge. Third, it may also diagnose the illness from the symptoms provided. 1.3.3 The Results 1. The main objective of this project called "Human ailment Prediction based totally on signs and symptoms" is to accurately diagnose the patient's illness by analyzing all current symptoms. 2. A graphical user interface is provided for the user to submit their symptoms in this assignment. 3. It takes the patient's data and symptoms into account while making sickness predictions. 14.4 Time-based Realization of the project's potential 1. This study will use datasets to forecast the patients' illnesses from their symptoms and other information. 2. Disease prediction allows for early detection of a patient's condition, allowing for the avoidance of potentially harmful illnesses. 3. To help people feel

more at ease utilizing the app instead of confiding in others about their health. 4. The qualities (symptoms) are presented with the required alternatives for selection.

Review of Literature

Current Systems Described Typically, this literature evaluation will focus on materials that bolster the examination of machine learning and methodologies relevant to hospitals. The research papers detailed all of the possible methods and frameworks used to categorize data. In their work, the authors used Decision Tree, Naive Bayes, Random Forest, and KNN as their machine learning algorithms. After an end handler inputs all the signs, all they have to do to get their result is click the Random Forest button. The results from all four algorithms are identical, according to this study. As previously stated, the accuracy was 0.976. The user's lack of algorithm knowledge is a major flaw in this system [1]. Using a variety of ML techniques, the authors developed a system for illness prediction. The analyzed data set included more than 230 illnesses. The breakdown method determines the illness as the output based on a person's symptoms, gender and age. Out of all the algorithms tested, the KNN algorithm has shown to be the most effective. The weighted KNN algorithm achieved an accuracy of 93.5%. Constant illness was a problem [2]. The illness forecasting process in this classification decision tree makes use of Random Forest, Naïve Bayes, SVM, and KNN algorithms. Accuracy of the system reaches 98.3%. Based on the algorithm's accuracy, we compared the algorithms' performance. off of all the algorithms, Decision Tree stands off with 84.5%, Random Forest with 98.95%, SVM with 96.49%, KNN with 71.28%, and Naïve Bayes with 89.4%. One main shortcoming of this system is that it needs extra time to execute due to the comparison of algorithm numbers [3]. The article delves into an algorithm that can diagnose a disease only by looking at its symptoms. The user interface includes a drop-down menu where the user may choose the symptoms associated with their ailment. The user's symptoms, in addition to the ailment, might inform the research's pharmaceutical recommendation [4]. A classification-based disease prediction machine for cardiovascular, renal, liver, diabetes, and breast cancer was developed by the authors. On average, we have a 95% chance of correct predictions. The effectiveness of the illnesses meant that the danger level was greater, relative to

other diseases [5]. The authors are making their 43 disease predictions based on a database of 106 characteristics that might be seen as symptoms. Following the discovery of the capacity to forecast

a disease. As a method, they used the KNN algorithm with Naïve Bayes. In order to foresee potential outcomes, this study incorporates the ML predictive method, which draws data from preexisting data units [6]. In their quest to effectively forecast potential illnesses, the authors are using machine learning methods such as SVM Support Vector Machine and MLR Multilinear Regression. For five distinct illnesses, they determined how well this technique worked. The suggested approach can provide results with an accuracy of 87% or higher. Since it can only forecast five illnesses, the diseases were severely constrained [7]. 2.2 Review of Relevant Literature: The prediction of illnesses is done using machine learning techniques. The frameworks employed are also cutting edge. This study's findings indicate the need for an approach with a shorter learning curve and an extensible foundation. It takes more time to train a version using previously suggested strategies. Furthermore, it provides less precise results. They tested a lot of algorithms side by side and picked the one with the highest accuracy for their final forecast.

PROPOSED METHOD

Approach to Design It is our hope that the development of sign-based disease prediction would help us combat common illnesses before they become catastrophic, hence reducing the number of premature deaths that occur as a result of people's general lack of knowledge about these potentially fatal conditions. All of the code for the "Disease Prediction based on signs and symptoms" project is written in Python. Even the assignment's user interface is implemented using T-kinter, a library interface in Python. With the help of a set of rules, the Random Forest algorithm and the Naive Bayes algorithm essentially complete this prediction. Before selecting symptoms from the drop-down box, the user must provide their name, age, gender, and blood group. The user may input their symptoms and then click the "predict" button to get a diagnosis of the condition that is associated with those indicators. The user inputs the symptoms they anticipate from the ailment; for more precise results, they should enter all of the symptoms. The computer then returns the proper result with an accuracy level between zero and one. A clear button is available to

reset the values, and an exit button is offered to quit the user interface or navigation. 3.2 Architecture of the System Diagram Machine learning-based disease prediction may make an educated guess as to whether or not a person has a certain illness based on their symptoms. A number of datasets that can analyze a person's symptoms and make predictions are a part of the machine learning-based Disease prediction Architecture's framework. Data is collected from multiple hospitals and individuals with diverse symptoms. In the preprocessing stage, any noisy data, missing values, or incomplete data may be cleaned. There are a number of methods for doing preprocessing; for example, if a tuple has a large number of missing values, it may be discarded or automatically filled in using mode. Other methods include binning, regression, and clustering. Manually filling in missing values is a tedious effort. Following the aforementioned steps, datasets are partitioned into smaller units using column-wise sorting. The data is then classified using classification algorithms. After that, the labelled statistics are processed into a pipeline that records undergo further processing before entering the disease prediction model. It takes in data in binary format and then converts it. In order to replace the provided symptom values with 1s, it will examine the 0s tuple of each illness. Then, it finds a tuple that is similar to or almost identical to all the entries in the dataset. Once the user inputs all of the following data, the system's prediction version will integrate and compare the processed information to provide a diagnosis of the disease. The building blocks, components, and guiding principles of an architecture may be visually represented in a structural diagram. Based on the gadget evaluation, the figure describes the system software. During training, a number of selection trees are constructed using an ensemble of machine learning approaches for tasks such as class and regression.

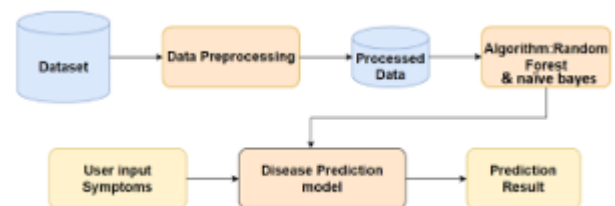


Fig 3.1 Architecture Diagram 3.3 Description of Algorithm

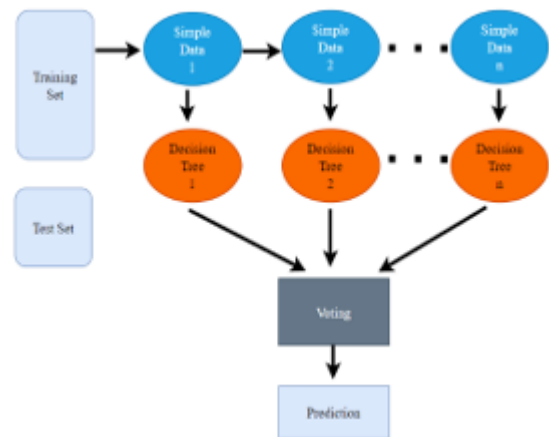
3.3.1 Algorithm for Random Forests Method of ensemble learning that constructs a large number of decision trees during the training process; used for problems such as classification and regression. Protocol 1: The RF Protocol [8]. Part of the Training: D) Train the dataset using the goal variable, p features, and ms occurrences. K is the number of target variable class labels. N: The quantum of RF classifiers. Process: From 1 to N, as m Step one is to generate a training dataset D_i sample at random. Second, for every random characteristic in every random sample, build a tree. Think about a node n (i) and choose q characteristics at random from the set $\sqrt{p}$ or $p/3$. (ii) Pick the optimal splitting technique before settling on a splitting feature.

Repeat this procedure for each additional node until the tree is complete. Proceed from step I to step III in the same order until all conditions are satisfied. 3. Build classifiers that have been trained C. Testing Phase: Merge the classifiers trained with N using a simple majority vote. Here is the projected Class label for test instance x, derived from the classifiers C:

$$C_B(x) = \underset{K(1)}{\operatorname{argmax}_j} \sum_{b=1}^N I(C_b(x) = j), \text{ for } j = 1, \dots$$

The Random Forest algorithm and how it works A first step is to choose random samples from a dataset and record their values. Step 2: Use the possibilities of the supplied talents to generate an opportunity table. Three, determine the posterior probability by using Bayes' theorem. Second Algorithm: Noob Bayes First, consider a training dataset D consisting of rows and the labels associated with them. A n-dimensional attribute vector $X = (x_1, x_2, \dots, x_n)$ represents each row in D. Consider a set of m classes denoted as $Y_1, Y_2, \dots, Y_m$. In order to find the largest posteriori, or the greatest $P(Q_i | A)$, classification is used. [2] in It is possible to infer this from Bayes' theorem (2). It is sufficient to maximize $P(X)$, as it remains constant across all classes. The second step is for these rules to construct a decision tree for each pattern. Decision trees for the chosen samples may be constructed using best-split-based metrics for attribute selection, such as information gain, gain ratio, and Gini index. Typically, the Gini index is used to determine the optimal splitting attribute. The prediction product from each selection tree will then be obtained. Thirdly, for each anticipated outcome,

vote casting may be carried out at this stage. The bulk of the outputs are regarded the final output for random forest after creating the decision tree. Fourth, take into account the most voted outcome when making your final forecast.



The subsequent diagram will illustrate its operating:

Pictured in Figure 3.2 is the Random Forest example. 3.3.2 Naïve Bayes: The Naïve Bayes algorithm takes into account the Bayes Theorem to do classification based on the likelihood of the class labels. To the uninitiated, "type algorithms" refer to programs that can sort fresh comments into existing lists of instructions. The New Bayes Classifier and How It Works: Initial Step: Make frequency tables from the provided dataset. Value counts are known as frequencies. The value's frequency is used to generate frequency tables.

in addition to its true worth. Step 2: Use the possibilities of the supplied talents to generate an opportunity table. Three, determine the posterior probability by using Bayes' theorem. Method 2: Naïve Bayes First, consider a training dataset D consisting of rows and the labels associated with them. A n-dimensional attribute vector $X = (x_1, x_2, \dots, x_n)$ represents each row in D. Consider a set of m classes denoted as $Y_1, Y_2, \dots, Y_m$. The maximum posteriori, or the maximum $P(Q_i | A)$ [2], are derived by classification. Theorem of Bayes may be used to deduce this.

$$P(Y_i | X) = \frac{P(X|Y_i)P(Y_i)}{P(X)} \quad (2)$$

$$P(Y_i | X) = P(X | Y_i) P(Y_i) \quad (3)$$

$$P(X|Y_i) = \prod_{k=1}^n P(x_k|Y_i) = P(x_1|Y_i) \times P(x_2|Y_i) \dots P(x_n|Y_i) \quad (4)$$

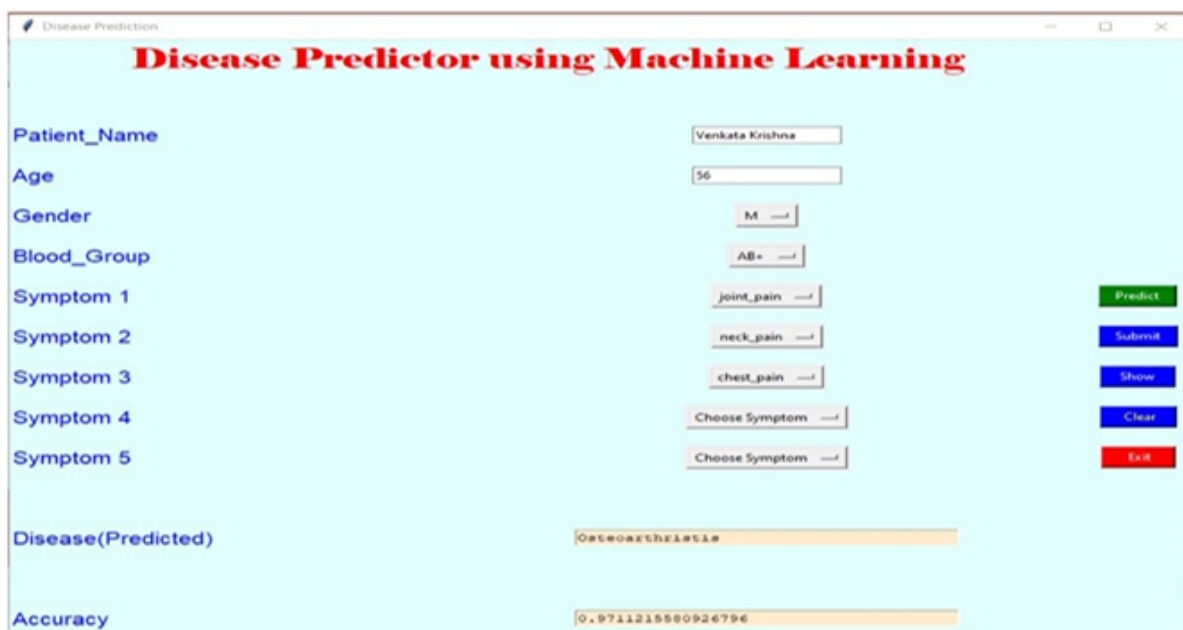
The Fourth Dataset Description Out of the 133 characteristics in the dataset, 132 are to symptoms and the remaining one is for prognosis. There is a binary format for the dataset. Put the ailments into numbers by multiplying each illness by a certain amount. The preprocessed dataset used in this study was obtained from GitHub. It is comprised of preprocessed patient records from different hospitals together with their associated illnesses and symptoms. The research dataset employed in this study provides the intended model with improved accuracy. There was a binary format for the dataset. Data should be converted to binary form after collection, and the model should be trained using the dataset.



Dataset

RESULTS AND OBSERVATION

Based on the five symptoms the patient reported, the GUI was able to accurately diagnose "Osteoarthritis" and save the patient's information for future reference (test case results). The illness may be predicted by looking at the



Disease Predictor using Machine Learning

Patient_Name: Venkata Krishna

Age: 56

Gender: M

Blood_Group: AB+

Symptom 1: joint_pain

Symptom 2: neck_pain

Symptom 3: chest_pain

Symptom 4: Choose Symptom

Symptom 5: Choose Symptom

Predict

Submit

Show

Clear

Exit

Disease(Predicted): Osteoarthritis

Accuracy: 0.9711215580924796

Fig 4.1 GUI Interface and Nb gives 96.88%.

	precision	recall	f1-score	support
accuracy			0.9752	1489
macro avg	0.9618	0.9729	0.9649	1489
weighted avg	0.9660	0.9752	0.9685	1489

Fig 4.2 Random Forest accuracy

	precision	recall	f1-score	support
accuracy			0.9711	1489
macro avg	0.9579	0.9688	0.9606	1489
weighted avg	0.9628	0.9711	0.9647	1489

Fig 4.3 Naïve Bayes accuracy

Given that the suggested approach compares Random Forest with Naïve Bayes. Both algorithms provide very similar results; rf achieves an accuracy of 97.29% at its maximum and an average of 96.7%. The maximum accuracy that naïve bayes may achieve is 97.11%, while its typical accuracy is 96.8%. If we look at accuracy, we see that rf achieves 96.18% and Nb 95.79%. The rf yields a recall value of 97.29 percent. 4.2 Findings from the project Taking five symptoms of a patient as a starting point for early disorder identification is the primary goal of this endeavor. A longer life for the sick person is possible as a result of early illness treatment. This illness model can identify the patient based on no more than five symptoms. The tool will anticipate the user's illness by collecting symptoms and comparing the naïve bayes and random forest algorithms' accuracies. It then finds the optimum method by comparing the recall and precision numbers.

CONCLUSION AND FUTURE WORK

In summary The goal of this machine learning project is to find a way to diagnose an illness only by looking at its symptoms. Two algorithms, Random Forest and Naive Bayes, are used. After the user inputs their symptoms and submits the form, the machine will utilize a comparison of the accuracies of the Random Forest and Naive Bayes algorithms to determine the

illness.

Looking Ahead This system is based on the GUI, so the user can access the system from anywhere in the world free of cost. In this way, one may identify the illness before seeing a doctor and take measures to stop it in its tracks. Some individuals, for instance, live in rural areas far from major cities, where access to modern medical care is limited; this would allow them to complete the required steps until they can go to a doctor. An opportunity to eradicate the illness at an early stage is created by employing this. Thanks to recent technological advancements, this equipment is now within everyone's grasp.

REFERENCES

- [1]. M Vamshi Krishna Reddy, G V P Sai Abhijith, K Sai Nath, Mangali Sathyanarayana(2022) Disease Predictor Based on Symptoms Using Machine Learning.
- [2]. Rinkal Keniya, Aman Khakharia, Vruddhi Shah,Vrushabh Gada, Ruchi Manjalkar, Tirth Thaker, Mahesh Warang, Ninad Mehendale.(2020) Disease Prediction from various Symptoms using Machine Learning .
- [3]. Dr C K Gomathy, Mr. A. Rohith Naidu (2021) The prediction of disease using machine learning
- [4]. Priyanka J. Panchal, Shaezrah A. Mhaskar, Tejal S. Ziman, "Disease Prediction using Machine Learning", 2020.
- [5]. Palle Pramod Reddy, Dirisinala Madhu Babu, Hardeep Kumar and Dr.Shivi Sharma, " Disease Prediction using Machine Learning ", 2021.
- [6]. Anjaiah. A, Sravya. V, Reddy. E. V, Vidya.K., Ankit. B. V. R., "Health Disease Prediction Based on Symptoms using Machine Learning", 2022.
- [7]. Md. Ehtisham Farooqui, Dr. Jameel Ahmad, Disease Prediction System using Support Vector Machine and Multilinear Regression ,2022.
- [8]. Smita Jaywantrao Ghorpade, Ratna Sadashiv Chaudhari, Seema Sajanrao Patil. "Enhancement of Imbalance Data Classification with Boosting Methods: An Experiment", ECS Transactions, 2022.
- [9]. Dhiraj Dahiawade, Prof. Gajanan Patle, Prof. Ektaa Meshram (2020) Designing Disease

Prediction Model Using Machine Learning Approach.

- [10]. Pahulpreet Singh Kohli, Shriya Arora (2022) Application of Machine Learning in Disease Prediction.
- [11]. Sneha Grampurohit and Chetan Sagarnal, "Disease Prediction using Machine Learning Algorithms", International Conference for Emerging Tehnology (INCET), 2020.
- [12]. Raj H. Chauhan, Daksh N. Naik, Rinal A. Halpati, Sagarkumar J. Patel and A.D. Prajapati, " Disease Prediction using Machine Learning", International Research Journal of Engineering and Technology, vol. 7, no. 5, pp. 2000-2002, 2020.
- [13]. M Maniruzzaman, M J Rahman, B Ahammed, M M Abedin Classification and prediction of diabetes disease using machine learning paradigmHealth Information Science and Systems, volume 8, issue 1 Posted: 2020
- [14]. G. Battineni, G. G. Sagaro, N. Chinatalapudi and F. Amenta, "Applications of machine learning predictive models in the chronic disease diagnosis", Journal of Personalized Medicine, vol. 10, no. 2, pp. 1-11, March 2020.
- [15]. [15] S. Uddin, A. Khan, M. E. Hossain and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction", BMC Medical Informatics and Decision Making, vol. 19, no. 281, pp. 1-16, December 2019.
- [16]. [16] D. Dahiwade, P. G. Patle and P. E. Meshram, "Designing disease prediction model using machine learning approach", 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211-1215, 2019.