Chikitsak: Disease Prediction system using Machine Learning

Mr. Prashant Kanade[1]
Assist. Prof., Dept. of Computer Engineering, VESIT
Mumbai, India

Mr. Amit V. Joshi[2]
Dept. of Computer Engineering, VESIT,
Mumbai, India

Mr. Puneet V. Meghrajani[3]
Dept. of Computer Engineering, VESIT,
Mumbai, India

Mr. Jayesh R. Shadi[4]
Dept. of Computer Engineering, VESIT,
Mumbai, India

***Abstract*—**Chikitsak is an intellectual prediction system which predicts an illness based on the information or symptoms entered into the system and provides the precise results. We propose a progressive alternative to the conventional method which resolves tedious problems of scheduling an appointment with the doctors. If one is not very serious and he/she only wants to know about the kind of disease he/she is facing this system is the cure for all ills. It is a system that provides the user with the tips and tricks to maintain the user's health system and provides a way to identify the disease through this prediction. Health industry plays an important role in the cure of patient's diseases, so it is also a kind of help to the healthcare industry that will inform the user and also help the user in the case he or she does not want to go to the hospital or to some other clinic, so that the user can get to know his/her condition by entering the symptoms and any other useful information.

Keywords-Chikitsak, Machine Learning, Disease Prediction, Symptoms, Healthcare, Django, Classification.

I. INTRODUCTION

Chikitsak is a system that predicts disease based on the information provided by the user. It also predicts the patient's or user's disease based on the information or symptoms entered into the system and provides accurate results based on that information. It is a system that provides the user the ease for maintaining the user's health, as well as a way to predict disease using the symptoms. Now a day's medical industry plays a big role in curing the patient's diseases as well as it also helps the healthcare industry. Therefore, this system provides the user with an alternative choice as if one does not want to go to a hospital or in any clinic, the user can just know the disease that he/she is suffering from, only by entering the symptoms and all other useful information. This DPUML (Disease Prediction using Machine Learning) has been done previously by many other organizations, but we intend to make it different and useful for the users who use this system. Today, doctors use numerous scientific technology and methods to identify and detect not only common illnesses, but many fatal diseases as well. A correct and accurate diagnosis is always the cause of the successful treatment. Doctors may not take accurate decisions when diagnosing a patient's disease, therefore systems that use machine learning algorithms help to obtain exact results in these instances. The disease prediction of patients using machine learning is designed to overcome general illness at an earlier stage. We all know there is a competitive environment in the healthcare industry but it also needs humanity and devotion towards their services to serve their purpose. According to research, 40 percent of the population does not worry about health. The main reason for the ignorance is that people are so concerned about their time and doctor, that they have no time to appoint and consult with their doctor, which is going to lead to fatal conditions later on. Research showed that 70% of the people in India have general illness and 25% are killed in early ignorance. The main reason for developing this project is that a user is able to sit down at their convenient location and have a health check, the UI is designed so simple that it can easily be used by everyone and checked.

II. LITERATURE SURVEY

A. Disease prediction by machine learning over big data from healthcare communities

With the rise of big data in the biomedical and healthcare communities, M. Chen, Y. Hao, K. Hwang and L. Wang [1]. suggested a solution in which reliable processing of medical data benefits early disease diagnosis, patient treatment, and the community resources. When the consistency of medical evidence is incomplete, however, the interpretation accuracy suffers. Furthermore, some regional infections have distinct symptoms in different countries, making disease outbreak prediction difficult.

The K-nearest neighbor algorithm is the machine learning algorithm used in this paper (KNN). This clearly demonstrates that a medical chatbot can diagnose patients with some accuracy using basic symptom diagnosis and a conversational approach using natural language processing.

B. Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning

A medical chatbot is designed to be a conversational agent that motivates users to address their health conditions and returns the diagnosis based on the symptoms presented by them in this method proposed by Rohit Binu Mathew, Sandra Varghese, Sera Elsa Joy, and Swanthana Susan Alex [2]. From user input, this chatbot device would be able to detect symptoms. The chatbot forecasts the condition and prescribes medication based on

the derived signs. Medical chatbots have a major influence on the state's health community. It has a higher level of dependability and is less vulnerable to human error. People stop going to the hospital with minor problems that might turn into a serious illness in the future. This problem is solved by the suggested solution. This concept revolves around developing a chatbot that is both free and accessible at all times of the day. The fact that the chatbot is free and can be used from anywhere, including the user's workplace, encourages them to have it and use it. It eliminates the costs of treating specialist doctors.

C. Disease Prediction using Machine Learning

The Disease Prediction method proposed by Kedar Pingale, Sushant Surwase, Vaibhav Kulkarni, Saurabh Sarage, Prof. Abhijeet Karve [5] was focused on predictive modelling, which forecasts the user's disease based on the symptoms they offer as feedback to the system. The machine analyses the signs presented by the patient as feedback and generates an appropriate output indicating the likelihood of the disease. The implementation of the Naive Bayes Classifier is used to predict disease. They have modeled diseases like Diabetes, Malaria, Jaundice, Dengue Fever, and Tuberculosis using linear regression and decision trees.

D. Big Data in Medical Applications andHealth Care.

Wang, L., and Alexander, C. A. [3] proposed Big Data will combine all health-related data and provide a 360-degree view of the patient, allowing analysts and forecasters to interpret and anticipate outcomes. It has the potential to transform clinical procedures, innovative drug growth, and the funding of health care. Early disease prevention, crime detection, and improved clinical delivery and reliability are only a few of the advantages. This paper discusses the Big Data definition and features, as well as health-care data and some of the more pressing Big Data questions. These topics cover the advantages of Big Data, its uses, and health care.

E. Prediction of heart disease using machinelearning algorithms

In this article by SanthanaKrishnan J, Geetha S [6], two supervised data mining algorithms were applied to a dataset to predict the likelihood of a patient developing heart disease, and the results were evaluated using the Naive Bayes Classifier and Decision tree classification models. These two algorithms are tested on the same dataset in order to determine which is the most accurate. The Decision tree model correctly predicted heart disease patients 91% of the time, and the Naive Bayes classifier correctly predicted heart disease patients 87% of the time. As a result, they conclude that the Decision Tree Classification algorithm is the easiest and most effective method for dealing with medical data sets. The developed framework, along with the machine learning classification algorithm, could be used to forecast or detect other diseases in the future. The work, which includes several other machine learning algorithms, can be generalized or developed for the automation of heart disease diagnosis. *Disease Diagnosis System by Exploring Machine Learning Algorithms*

In recent years, Allen Daniel Sunny, Sajal Kulshreshtha, Satyam Singh, Srinabh, Mohan Ba and H Sarojadevi [7] have seen a troubling pattern sweeping the developing world. Medical costs are skyrocketing in countries that are known around the globe for being technological and scientific trailblazers. The United States and western European countries are among the countries. Health care costs are increasing due to a variety of reasons, including very high prescription medication prices and the use of more expensive medical procedures. Rising medical costs seem to be linked to a country's overall growth. Thus, they concluded that algorithms like Naive Bayes and Apriori are extremely useful for disease diagnosis on the provided data set based on the different algorithm implementations. We should assume that all objectives have been reached depending on the task, with the illness being forecast using various algorithms based on the input symptoms. Since machine learning algorithms rely on datasets based on prior information gained from physical diagnosis, the dataset volume must be large and contain a small number of outliers. If more diseases are added to the database, the spectrum of diagnosis expands, allowing for improved future forecasts.

III. METHODOLOGY

A. Naive Bayes:

In data mining, naive Bayes classifiers are a subset of basic probabilistic classifiers based on Bayes theorem and solid part autonomy assumptions. It was first applied to the text rescue community in the early 1960s under a different name and has since become a common tool for text categorization, solving the issue of evaluating documents belonging to only one of two categories using word frequencies of the elements. It became successful in this domain with highly sophisticated methods such as bolster vector machines after proper preprocessing. It also finds use in computer-assisted medical diagnosis. Another category is the sceptic. Bayes classifiers are modular, requiring a number of parameters that are proportional to the number of variables in a learning problem. Instead of an expensive iterative guess, as is used with certain other types of classifiers, the best likelihood preparation should be possible by evaluating a closed frame expression, which takes straight time. Naive Bayes models are recognized by a variety of names in the measurements and software engineering literature, including clear Bayes and autonomy Bayes. While both of these titles allude to the use of Bayes' theorem in the classifier's decision-making process, innocent Bayes is not a Bayesian strategy.

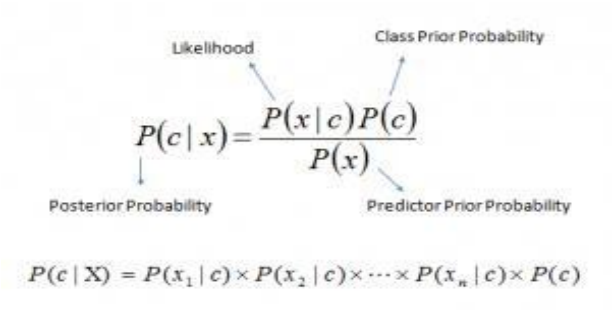

Fig 1. Naive Bayes Theorem

B. Decision Tree:

A decision tree is a flowchart structure in which each internal node represents a test on a particular attribute, each branch represents the test's outcome, and each leaf node represents a class name. The grouping laws are shown by the pathways from the root to the leaf. The interconnected diagram is used as an empirical, visual, and decision-making instrument in tree, where the visible values are measured. A decision tree has three kinds of nodes: decision nodes (represented by squares), chance nodes (represented by circles), and end nodes (represented by triangles). A decision tree begins with a choice that must be made. To display the tree on the left side of a large sheet of paper, draw a small rectangle. Draw lines in the right-hand direction for each possible solution, then write the solution along the graph. Represent the effects at the end of each section. If the decision's outcome isn't clear, draw a wide circle. If the outcome is an alternative decision, you would need to draw a new rectangle. Squares signify choices, while circles represent an unknown outcome. Draw lines denoting alternatives that could be selected starting from the new decision squares from the diagram. We can also draw lines that describe potential outcomes from the circles. Finally, jot down a quick note on the line by repeating what it says.

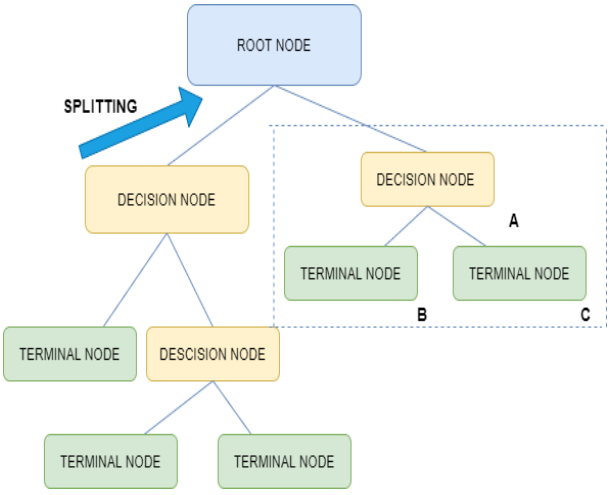

Fig 2. Decision Tree

C. Random Forest:

Random forest is a cart-based bootstrapping algorithm. It constructed several trees with various initial variables, taking into account a sample of 100 observations and five randomly selected initial variables to construct a cart model. The procedure will be repeated ten times, and they will make a final prediction for each observation. Each forecast affects the final prediction. This last one will easily be the average of all the predictions. Essentially, the Weka tool is used to complete this operation. Weka is a machine learning platform that includes a wide number of data science algorithms that can be used for classification, estimation, and missing value detection. It's a group learning tool for organization and other activities that works by assembling a collection of decision trees during training and generating a class that's the mode of the classes or the mean predictor of the single trees. Tin Kam shaped the mechanism for the random woods. In Ho's formulation, the random subspace method is a process for implementing Eugene Kleinberg's "stochastic discrimination" method to classification. The method outlined above is the initial tree bagging algorithm. Randomforests differ from this general arrangement in one way: they used a revamped tree learning algorithm that selects a random subset of features for each individual in the learning process. "Feature bagging" is a term used to describe this operation. The numerous trees in a typical bootstrap study emphasize the significance of this correlation: if one or more additional features are accurate predictors for the answer variable, these features will be sensibly selected in several of the B trees, resulting in correlation

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2$$

Where N is the number of data points,
fi is the value returned by the model and
yi is the actual value for data point i.

Fig 3. Random Forest

IV. CONCEPTUAL ARCHITECTURE

Machine learning disease prediction forecasts the user's existence of the disease on the basis of different symptoms, and the information provided to them by the user such as headache, back pain, runny nose and much more about the symptoms. Machine learning consists of a variety of datasets to equate the user's symptoms and simulate the configuration of the device disease detection, the data sets

shall then be converted into smaller sets and then sorted according to classification algorithms and subsequently transformed into machine learning systems from which the data is being processed, using all the user inputs indicated above, and entering the disease prediction model. The patient then integrates and checks on the prediction model of the system after entering the details and the data processed as a whole and forecasts the disease.

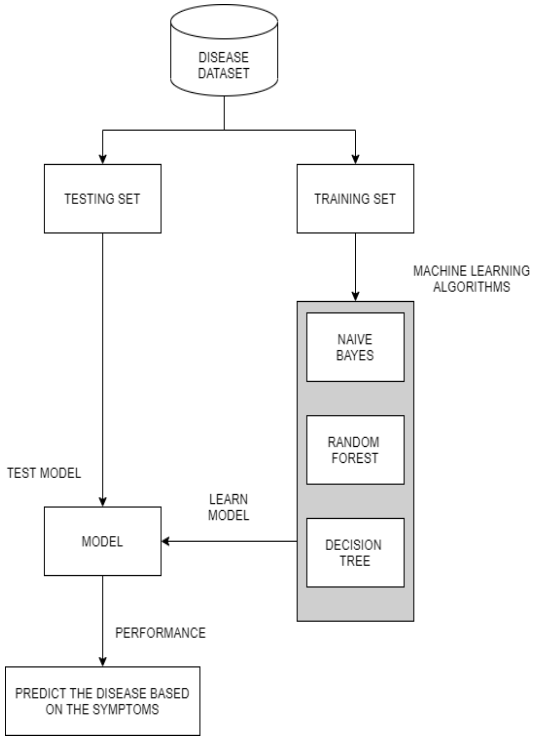

Fig 4. Conceptual Architecture of the System.

1. The train-test split procedure is used to estimate the performance of machine learning algorithms. Our dataset consists of symptoms and their corresponding diseases.
2. The dataset is split into training set and testing set using sklearn library. Once the data is preprocessed and split into training set different classification algorithms are tested over the data.
3. Algorithms such as Naive Bayes, Random Forest, Decision Tree gave us the most accurate results.
4. Finally the disease is predicted.

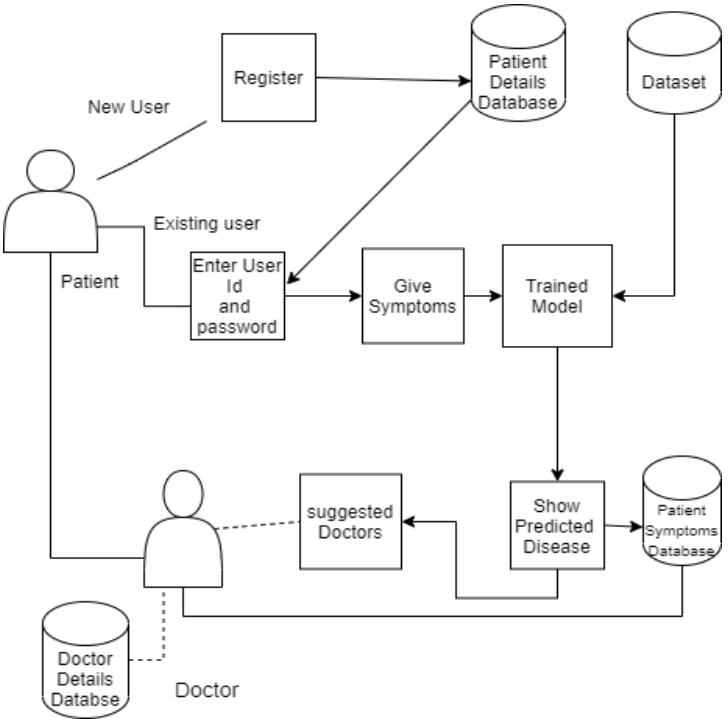

Fig 5. Block Diagram of the System

1. The disease prediction method has three users: a doctor, a patient, and an administrator.
2. The system authenticates each of the system's users.
3. The system has a role-based access system.
4. The machine allows the patient to enter the symptoms, and the system then predicts an illness based on those symptoms.
5. For diseases that are predicted, the system recommends the corresponding doctors.
6. The system allows online consultation for patients.
7. The system allows patients to consult with doctors from the comfort of their own homes.

V. IMPLEMENTATION SETUP

Patients and doctors can interact with this system using our GUI created using Django. Patients after signing in will be prompted to a screen where they can see their profile, they can enter the symptoms faced and can get the disease from the trained model and can consult a specialist doctor and can chat with them. They can check consultation history and can give feedback. For doctors sign in, there is a list displayed of the patients . Doctor can consult each patient where he can see the predicted disease with symptoms and can get back to the patient with the diagnosis of the following disease .

VI. RESULT AND ANALYSIS

A. Setup Environment

1) All of the experimental cases are written in Python with Django and PostgreSQL as backend, algorithms and methods, and contrasting classification approaches, as

well as many features extraction technique, and run on a system with an Intel Core i5-6200U, 2.30 GHz Windows 10 (64 bit) processor and 8GB of RAM.

B. Result Comparison

Algorithm	Accuracy
Naive Bayes	92.857
Random Forest	93.462
Decision Tree	97.619

Fig 6. Accuracy of Different Algorithms

Now, once the accuracy of the different algorithms is achieved, we have to implement the best algorithm in our prediction model. The user will enter the symptoms he/she is facing and the trained model will predict the disease. The chat box in consult a doctor section helps the system to keep an interactive environment for the user through which the diagnosis of the patient can be done earlier without any delay.

Inputs (Symptoms)	Output (Diseases)
Fatigue, Constipation, Weight gain, Muscle weakness, Puffy face.	Hypothyroidism
Fatigue, blurred and distorted vision, excessive hunger, continuous feel of urine.	Diabetes
Fatigue, high fever, loss of appetite, malaise, sweating	Tuberculosis
Chills, fatigue, high fever, nausea, vomiting, loss of appetite	Dengue

Fig 7. Disease Prediction Table

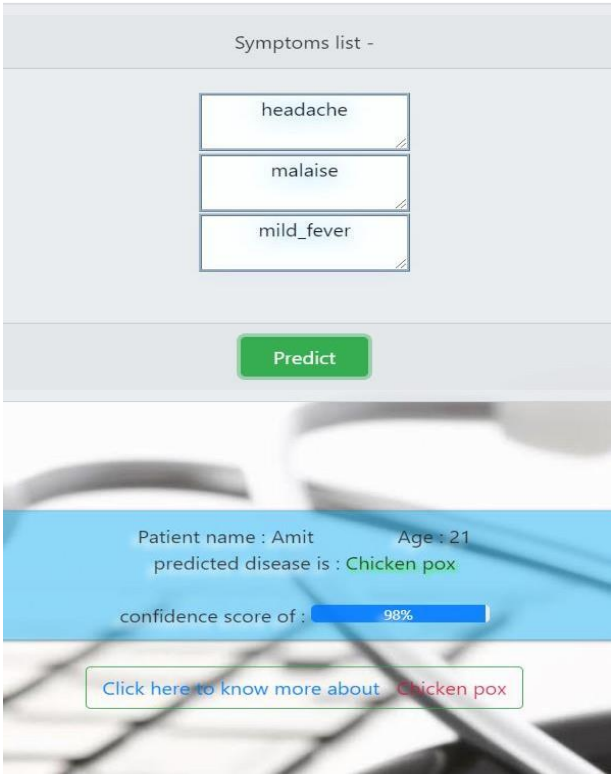

Fig 8. Prediction of disease

Thus, the disease is predicted by the trained model and shown on the screen. The confidence score is the probability that disease has occurred from the given symptoms. After the prediction the system provides a google link to know more about the disease. The system also categorizes the disease in different specializations. for e.g. - chicken pox comes under the specialization dermatologist. The system suggests the doctors which are registered with the system with their specialization. The patient can consult the doctor and can live chat with the doctor. The doctor can see the symptoms of the patient and can also see the disease predicted and the doctor can consult accordingly. This provides the two-way interaction between doctor and patient.

VII. CONCLUSION

Finally, we would like to state that this project Disease prediction using machine learning is extremely useful in everyone's day-to-day lives, but it is especially valuable for the healthcare sector, as they are the ones who use these systems on a daily basis to predict the diseases of patients based on their general information and symptoms. Now that the health industry plays such an important role in treating patients' illnesses, this is a valuable tool for the health industry to inform the consumer, as well as for the user if he or she does not want to go to the hospital or some other clinic. So, by simply entering the symptoms and all other relevant details, the user can learn about the disease from which he or she is suffering, and the health industry can benefit from this method by simply asking the user for symptoms and entering them into the system, where they can learn about the disease in a matter of seconds. If the health industry adopts this project, doctors' workload will

be minimized, and they will be able to foresee the patient's illness more easily.

VIII. FUTURE SCOPE

The method still requires the doctor's knowledge and experience due to alternative factors starting from medical records to weather conditions, atmosphere, blood pressure and numerous alternative factors. Limited number of diseases are currently present in the dataset. The scope of the dataset can be improved. Analysis of data, cleaning of data and delivery of results are slow and takes time. Past history of the disease has not been considered. This is a main factor on which more work has to be done in future. The Prediction is completely based on symptoms, factors such as age, climate are not taken into consideration. Existing system can predict the disease but not the subtype of the disease. Some local diseases should be explicitly added which can improve the scalability of this system. A centralized system can also be designed by allotting unique id numbers to the users for better handling of data. Generic medicine can also be displayed as an alternative for the medicine recommended by the doctor. The facility of uploading the reports of the tests recommended by the doctors can be added to improve consultation experience.

IX. REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities"," IEEE Access,vol. 5, no. 1, pp. 8869–8879, 2017.

[2] Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning -Rohit Binu Mathew, Sandra Varghese, Sera Elsa Joy, Swanthana Susan Alex

[3] Wang, L., and Alexander, C. A., Big Data in Medical Applications and Health Care. Current Research in Medicine 6:1–8, 2015.

[4] Abdulsalam Yassine, S., Mining Human Activity Patterns from Smart Home Big Data for Health Care Applications. IEEE Access 5:13131–13149, 2017.

[5] Disease Prediction using Machine Learning Kedar Pingale, Sushant Surwase, Vaibhav Kulkarni, Saurabh Sarage, Prof. Abhijeet Karve

[6] SanthanaKrishnan J, Geetha S (2019) Prediction of heart disease using machine learning **algorithms**. In: 2019 1st international conference on innovations in information and communication technology (ICIICT)

[7] Allen Daniel Sunny, Sajal Kulshreshtha, Satyam Singh, Srinabh, Mohan Ba and H Sarojadevi, "Disease Diagnosis System by Exploring Machine Learning Algorithms'', International Journal of Innovations in Engineering and Technology (IJIET), vol. 10, no. 2, May 2018.

[8] Ashok Kumar Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease" in Computer Applications and Mathematics, © Springer, 2016.

[9] M.A. Jabbar, B.L. Deekshatulu and Priti Chandra, "Intelligent heart disease prediction system using random forest and evolutionary approach", Journal of Network and Innovative Computing, vol. 4, pp. 174-184, 2016.

[10] S.Leoni Sharmila, C.Dharuman and P.Venkatesan "Disease Classification Using Machine Learning Algorithms - A Comparative Study", International Journal of Pure and Applied Mathematics Volume 114 No. 6 2017, 1-10