Multiple Disease Prognostication Based On Symptoms Using Machine Learning Techniques

Kajal Patil[1,*], *Sakshee* Pawar[2], *Pramita* Sandhyan[3] and *Jyoti* Kundale[4]

[1,2,3] Ramrao Adik Institute of Technology, Nerul, Navi Mumbai, India

[4] D.Y. Patil Deemed to be University, Ramrao Adik Institute of Technology, Nerul, Navi Mumbai, India

Abstract. Disease Prediction system that uses Machine Learning forecasts the ailments on the basis of the data pertaining to the symptoms entered by the user and provides trustworthy findings based on that data. If the patient isn't in any danger and the user merely wants to know what kind of ailment he or she has had. It is a system that gives the user suggestions and methods on how to keep their health system in good shape, as well as a way to find out if they have a sickness utilizing this forecast. Due to a diversity of diseases and a lower doctor-patient ratio, the use of particular disease prediction technologies as well as concerns about health has risen. We are focusing on offering customers with an instant and accurate disease prognosis based on the symptoms they enter, as well as the severity of the condition projected. It will provide the best algorithm and doctor consultation. Different machine learning algorithms are employed to forecast illnesses, ensuring speedy and reliable predictions.

1 Introduction

Medical science and universal health care are critical components of the economy and human existence. There has been a significant amount of shift between the world we live in now and the world that existed only a few weeks ago. Everything has become dull and erratic.Today's demand for medical industries have a greater requirement for data mining. When certain data mining techniques are applied correctly, important information may be mined from a huge database, which can aid medical practitioners in making early judgments and improving health care. The goal is to aid the physician by using the categorization.

Machine learning has become more popular as technology has advanced, owing to greater computer power and the availability of datasets on open-source sources. Machine learning is used in a number of ways in health care. The health care industry generates a vast quantity of data in the form of photographs, patient data, and other types of information that may be used to spot trends and make predictions. In health care, machine learning is utilize to tackle a variety of issues. Individuals with heart illness have varying degrees of heart disease, and the severity of heart disease varies from person to person.

As a result, creating a machine learning model, training it on the data set, and incorporating unique patient information can aid in prediction. The forecast result will be based on the information provided, and hence will be unique to that person. Type-2 diabetes is a condition that may be avoided by maintaining a healthy weight, lifestyle, and other factors. Corona virus is a condition for which there is no apparent cure. COVID-19 is a corona virus that emerged in China. This Disease Prediction using Machine Learning is totally done with the aid of Machine Learning, and we will forecast the disease using the data set that was previously provided by the hospitals. Numerous scientific methods and methodologies are being used by doctors for the identification and diagnosis of not only common ailments, but also many deadly disorders.

The proper and precise diagnosis is always credited with a successful therapy. Doctors may occasionally make incorrect judgments while diagnosing a patient's sickness; as a result, disease prediction systems that employ machine learning algorithms aid in obtaining accurate findings in such circumstances. It was created to combat general disease at an earlier stage, as we all know, in the competitive environment of economic development, mankind has become so engrossed that he or she is no longer concerned about health. It can detect people who are at risk of sickness or other health problems.

Clinicians can then take the necessary precautions to avoid or reduce the risk, so improving the quality of care and avoiding unnecessary hospitalizations. Disease risk prediction may now take advantage of vast volumes of semantic data, such as demographics, clinical diagnosis and measures, health behavior, test findings, prescriptions, and care mutilation, thanks to recent advancements in data analytic tools and methodologies. Electronic health data might be a good option for constructing illness prediction models in this case. Over time, a great number of illness prediction models based on large-scale electronic health databases have been proposed in the literature. Stakeholders such as the government and health insurance firms may gain from disease prediction. It can detect people who are at risk of sickness or other health problems.

* Corresponding author: 2906kajal@gmail.com

1.1 Purpose and Problem Definition

If a person has been observing a few symptoms and is unsure of the sickness he or she is dealing with, this will lead to a variety of diseases in the future. To avoid this and to learn about the condition in the early phases of the symptoms, this disease prediction will be extremely beneficial to a wide range of individuals, including children, teens, adults, and elderly citizens. Preventive measures can be taken by the individual or can seek for professional medical advice in order to get the proper treatment required to treat the ailment.

The scope of a disease prediction system is specifically vast, conferring how the world is evolving and along with the advancements in the technology comes with a lot of cons such as the various adulterates of food items, lack of proper nutrient supply to the body, unhealthy lifestyles involving improper food consumption as well as problems like obesity or unhealthy weight. All this accompanies with a variety of diseases.

Unfortunately as people are too engrossed in their daily activities they neglect their health. Children and senior citizens can also ignore or fail to recognition the important symptoms which can result in greater issues later on. It is wise to get treated before the disease develops and grows more detrimental. A prediction system can help such people to detect and act upon their health issues at an early stage and take preventive care. This also helps in getting primary health care in remote areas.

2 Literature Survey

Numerous research works have been carried out for the prediction of the diseases based on the symptoms shown by an individual using machine learning algorithms. Mont et al. [1] designed a statistical model to predict whether a patient had inuenza or not. They included 3744 unvaccinated adults and adolescent patients of inuenza who had fever and atleast 2 other symptoms of inuenza. Out of 3744, 2470 were confirmed to have inuenza by the laboratory. Sreevalli et al. [2] used the random forest machine-learning algorithm to predict the disease based on the symptoms.

The system resulted in low time consumption and minimal cost for the prediction of diseases, algorithm resulted in an accuracy of 84.2%. [4] considered medical records of 4920 patients and identified 132 symptoms corresponding to 41 diseases and implemented this dataset by using Naïve Bayes, Decision Tree and Random Forest algorithms. In this system, Naïve Bayes showed the highest security comparatively of 93.61%. The researchers in [5] used Convolutional Neural Networks and K-Nearest Neighbour algorithms are implemented on Disease dataset from UCI machine repository.

The results found through comparative analysis is that the CNN performed better than KNN in accuracy and time comparison. [6] used UCI heart disease repository and implemented Support Vector Machine to classify attributes into different heart disease classes. In this, SVM achieved an accuracy of 89%. In [7], the associated diseases with the symptoms using Random Forest where UCI disease repository is used. In the result, Random Forest achieved an accuracy of 85%.

3 Proposal

3.1. Module Description

Methods of machine learning have been available for a long time, and they've been compared and used in a number of ways for data science analysis. The major purpose of this research project was to examine the strategies for selecting features, preparing data, and processing data in machine learning training models. The challenge we have today with first-hand models and libraries is data, which, in addition to their amount and our cooked models, has a wider variety in the accuracy we see throughout training, testing, and real validation.

Furthermore, because the goal of machine learning is to produce an appropriate computer-based system and decision support that may help in the early diagnosis of diseases, we constructed a model in this project that uses machine learning algorithms to classify whether a patient will have any ailment. As a result, early illness prediction can assist in making decisions about lifestyle adjustments in high risk patients, reducing consequences, which can be a significant milestone in the area of medicine.

In order to look into the verified diseases and their corresponding diseases, an authentic data set was collected which had the list of the diseases and their symptoms listed in the database. An automated technique was used to create a knowledge database of disease-symptom connections based on information from textual discharge summaries of patients hospital to New York Presbyterian Hospital in 2004.

The disease is listed first, followed by the number of discharge summaries with a positive and current mention of the condition, as well as the accompanying symptom. Relevant algorithms were studied and compared with the help of the pre-exist literature and the results were observed. Accordingly, three algorithms were found to be suitable and then compared with the results generated by them individually. The selected algorithms were Naive Bayes, Decision tree and Random Forest.

Given the list of symptoms a patient is currently showing, our proposed system shall predict the likely diseases using an appropriate machine learning algorithms. Based on this data, we shall apply a set of machine learning

algorithms to comparatively evaluate which algorithm performs better among them. The algorithms used are Multinational Naive Bayes Algorithm, , Decision Tree Algorithm, Random Forest Algorithm and which will help us in getting accurate predictions.The data set will be derived from many open source assets available on the research groups like John Hop kin university and Columbia University.

We can detect the chronic kinds of disease in a certain location and population using structured analysis. With the aid of algorithms and methodologies, we choose the characteristics automatically in unstructured analysis. Based on the user's symptoms and prior assets, this algorithm forecasts the condition. It also assists in the constant examination of viral illnesses, heart rate, blood pressure, sugar level, and other information stored in the system, and it forecasts the suitable and correct disease based on those symptoms as well as other exterior symptoms.The system captures the user's symptoms and uses these, as well as prior assets, to forecast the condition. It also aids in the constant examination of viral illnesses, heart rate, blood pressure, sugar level, and other information stored in the system, and it forecasts the proper and correct disease based on those symptoms as well as other exterior symptoms.

Step 1:

Data collection and data set preparation The act of gathering, cleaning, and combining data into a single file or data table, particularly for use in analysis, is known as data preparation.

Step 2:

Developing a probabilistic modeling and machine learning approach Probabilistic Models in Machine Learning is the use of statistical coding to data analysis. The use of probabilistic models to define the world is portrayed as a common idiom.

Step 3:

Training and experimentation on database The first data needed to train machine learning models is known as training data (or a training data set). Machine learning algorithms are taught how to make predictions or perform a task using training assets.

Step 4:

Deployment and analysis Deployment is the process of integrating a machine learning model into an existing production environment in order to make data-driven business decisions. It's one of the last steps in the machine learning process, and it's also one of the most time-consuming.

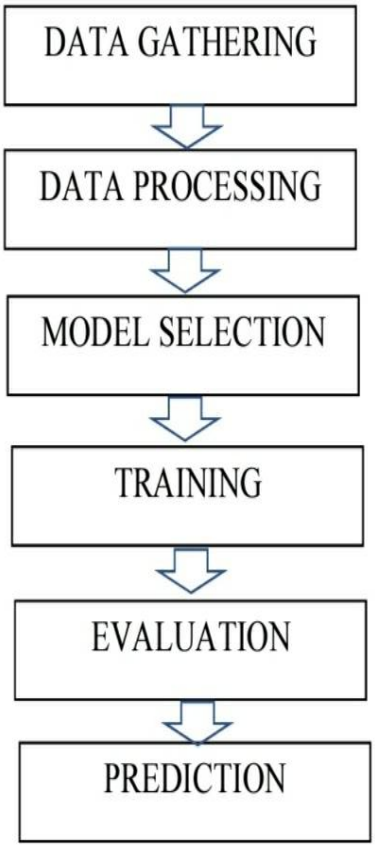

Fig. 1. Contraflow diagram

3.2 Hardware and Software Requirements

Hardware Requirement:

- 4 GB RAM.
- 256 GB HDD.
- Intel 2.8 GHz i3 Processor

Software Requirement:

- Windows 7/8/10
- Python
- Jupyter notebook.

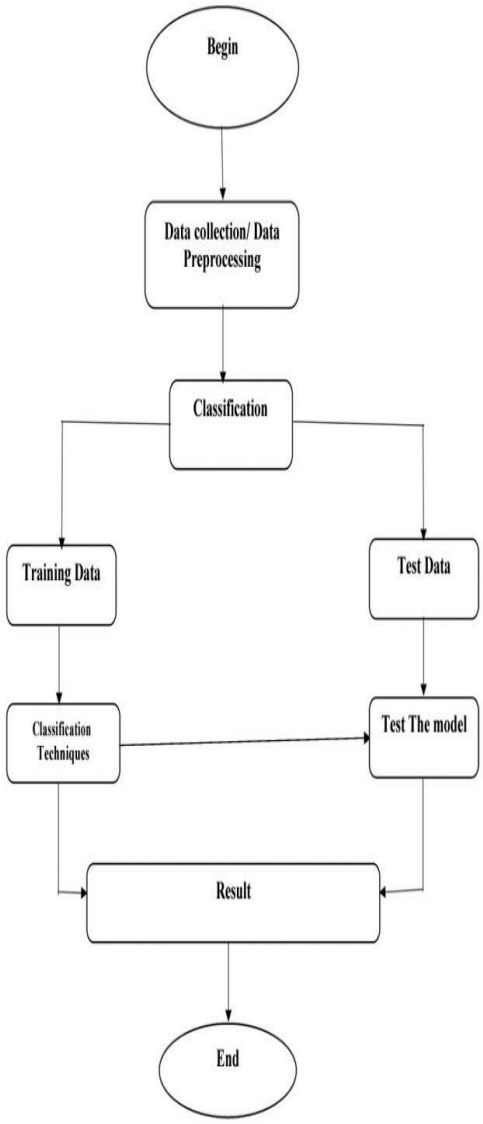

Fig. 2. Work flow of the proposed system

The initial step is to collect the relevant data and process it using the advanced data cleansing techniques. Later, the data set is divided into two divisions, training data and testing data so as to prepare the machine learning model effectively. After the ML model is ready, it is tested using a variety of combinations of symptoms and the result is verified.

4 Experimental Results

4.1 Result and Analysis

Multinomial Naive Bayes

It's a tribe of algorithms and not one algorithm. All naive socio-economic classifiers in Bayes assume that, given the class variable, the value of a certain feature is independent of the value for each other. The Thomas Bayes classifier might be the straight probabilistic classifier that supports the use of theorems with strong naive assumptions of independence (from Bayesian statistics). The Bayes theorem with associated grades of independence between the predictors was supported by classification method. Simply put, the presence of a special feature in an extremely class is assumed by a Naive Thomas Bayes category to be distinct from that of the other.

For example, if this is red, spherical, and around a few inches' diameter, the fruit may also be considered as an association grade apple. Even though those characteristics are interdependent and dependent on other characteristics, a naive classifier in Bayes would regard all these attributes separately to contribute to the likelihood that this fruit is an apple. The class label for each training data set is predicted by these learners. The class label which most of the models predict is voted using the majority voting procedure and the class label is agreed on in the training data set. Regulations are produced from the ensemble models.

The Multinomial Naive Bayes method is a probabilistic learning technique commonly used in Natural Language Processing (NLP). Using the Bayes principle, the algorithm estimates the tag of a text such as an email or a newspaper article. It assesses the likelihood of each tag for each sample and outputs the tag with the highest likelihood.

The Naive Bayes classifier is made up of several algorithms that all have one thing in common: each feature being classed is unrelated to any other feature. The presence or absence of one trait has no bearing on the bother's existence or absence. Accuracy achieved on our dataset: 0.92

Decision Tree

Building classification models using decision tree induction is a non-parametric technique. The challenge of finding the best decision tree is NP-complete. Decision tree construction techniques have been designed to be computationally cheap, allowing models to be built even when the training set size is quite high.

Decision trees, particularly those of a smaller size, are quite simple to read. For learning discrete-valued functions, decision trees give an expressive representation. The presence of noise is not a problem for decision tree algorithms, especially when approaches for diminishing over-fitting are used.

Decision tree induction is a non-parametric approach for creating classification models. Although constructing an optimal decision tree is an NP-complete problem, computationally affordable ways for creating decision trees have been identified, allowing models to be generated even when the training set size is relatively large.

Decision trees, especially ones of a smaller size, are easy to understand. Decision trees provide an expressive model for learning discrete-valued functions. Noise is not a concern for decision tree algorithms, especially when over-fitting minimization techniques are applied. On our data set, we attained an accuracy of 0.97.

Logistic Regression

Logistic Regression is a Machine Learning technique for resolving categorization problems. It's a type of predictive analytic approach based on the probability concept. To forecast the probability of a categorical dependent variable, the classification algorithm Logistic Regression is utilized. In logistic regression, the dependent variable is a binary variable with data coded as 1 (yes, True, normal, success, etc.) or 0 (no, False, abnormal, failure, etc).

The purpose of Logistic Regression is to find a relationship between a set of attributes and the probability of a particular event. A Logistic Regression model is similar to a Linear Regression model, except that instead of using a linear function, it uses a more complicated cost function known as the "Sigmoid function." The accuracy achieved on our data set: 0.89

4.2 Dataset

It is a set of disease-symptom connections developed by an integrated strategy using data from written clinical notes of admitted patients to New York Presbyterian Hospital in 2004. Depending on these notes, the most common illnesses were calculated, and the symptoms were graded based on the strength of relationship. Every sickness has a clinical manifestations that can be used to forecast the condition.

Table 1. Sample dataset with symptoms with the diseases

Symptoms	Disease
Runny nose, sore throat, cough, congestion, body aches, headaches, sneezing, fever	Common cold
Fever, profuse sweating, headache, nausea, vomitting, diarrhoea, anemia, muscle pain, convulsions	Malaria
Poor appetite, abdominal pain, headaches, generalized aches, pains, fever, lethargy, intestinal bleeding or perforation, constipation	Typhoid

In the same manner as shown in the above table, the symptoms are related with the associated disease. That is, the similar set of symptoms are grouped together and linked with a particular disease to indicate that the individual must be suffering from that disease if the symptoms experienced are matching with the dataset. By this method, the disease is predicted based on the symptoms which are taken as the input from the patient. The model is equipped to predict a plethora of diseases from the combination of symptoms, the diseases that can be predicted are fungal infection, allergy, peptic ulcer disease, diabetes, gastroenteritis, bronchial asthma, hypertension, jaundice, malaria, chickenpox, typhoid, dengue, tuberculosis, common cold, pneumonia are some of them.

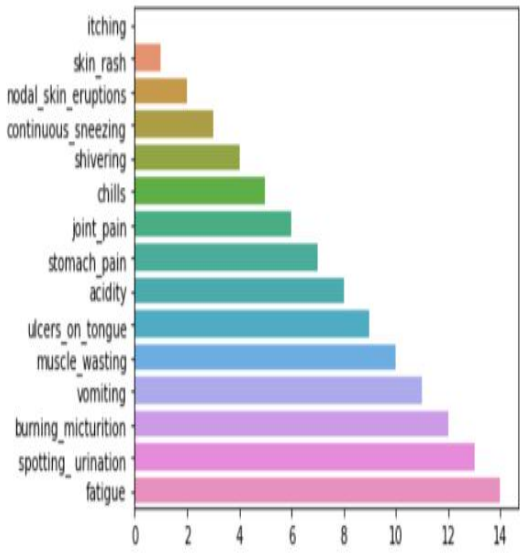

Fig. 3. Occurrence of the symptoms in the test data set

The above figure demonstrates the frequency of the symptoms occuring in the data set and shows which symptom is associated with the maximum of the diseases. The data set contains various diseases and the symptoms associated with them. The model predicts the disease which matches with the symptoms in the data set with the set of symptoms entered by the user. With this methodology, we can predict the disease which is most likely associated with the set of symptoms already present in the data set.

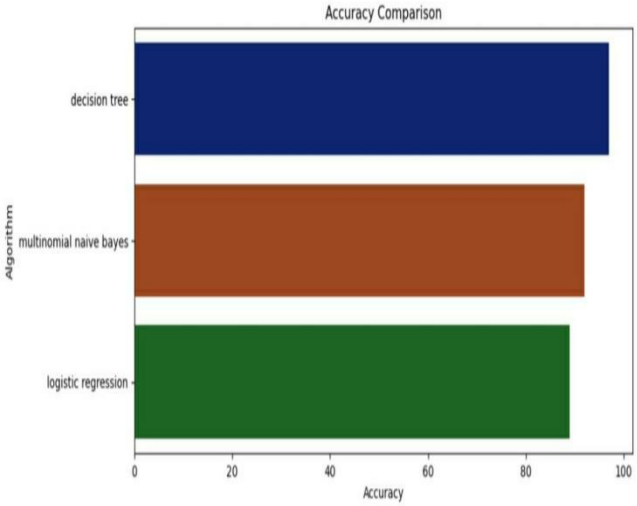

Fig. 4. Accuracy comparison of the algorithms

The accuracy comparison shown in the Fig. 4. portray the comparison between the three algorithms - Decision

Tree, Multinomial Naive Bayes and Logistic Regression plotted against a scale of accuracy percentage. The data set which was collected from an authorized source was then made suitable for use of the machine learning algorithms by data cleaning. The model was developed for Multinomial Naive Bayes algorithm, Decision Tree algorithm and Logistic Regression algorithm. The accuracy of Multinomial Naive Bayes on the chosen data set was 92%. As for Logistic Regression, it was merely 89%.The highest accuracy among all these algorithms was demonstrated by Decision Tree which was 97%.

5 Conclusion

Because medical data is growing at an exponential rate, it is necessary to process existing data in order to predict precise disease based on symptoms. Several general disease prediction systems based on machine learning algorithms such as Decision tree, Logistic Regression, and Multinomial Naive Bayes have been proposed to classify patient data. Sickness and risk prediction may be done in a short amount of time and at a minimal cost using this method. To summarise, people who are always concerned about their health will benefit from our method. Our objective is to develop this system so that people are more aware of their health issues and may lead a healthier lifestyle. Furthermore, depending on the user's symptoms, our system will provide more accurate sickness forecasts, as well as inspiring thoughts and graphics. As a consequence, we can say that our system has no limitations because it can be used by everyone.

References

1. A.S. Monto, S. Gravenstein, M. Elliott, M. Colopy, J. Schweinle, Clinical signs and symptoms predicting inuenza infection, Archives of internal medicine 160(21), 3243 (2000)
2. B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, “A relative similarity based method for interactive patient risk prediction,” Springer Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070– 1093, 2015.
3. D.R. Langbehn, R.R. Brinkman, D. Falush, J.S. Paulsen,M. Hayden, an International Huntington’s Disease Collaborative Group, A new model for prediction of the age of onset and penetrance for huntington’s disease based on cag length, Clinical genetics 65(4), 267 (2004)
4. S. Grampurohit and C. Sagarnal, ”Disease Prediction using Machine Learning Algorithms,”2020 International Conference for Emerging Technology(INCET),2020, pp. 17, doi: 10.1109/INCET49848.2020.9154130.
5. D. Dahiwade, G. Patle and E. Meshram, ”Designing Disease Prediction Model Using Machine Learning Approach,” 2019 3rd International Con- ference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.
6. Automatic Heart Disease Prediction Using Feature Selection And Data Mining Technique Le Ming Hung,a, Tran Ding,Journal of Computer Science and Cybernetics, V.34, N.1 (2018), 3347 DOI: 10.15625/1813- 9663/34/1/12665
7. International Journal of Scientific Research in Computer Science, E., & IJSRCSEIT, I. T. (2019). Generic Disease Prediction using Symptoms with Supervised Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. https://doi.org/10.32628/CSEIT1952297.
8. Pingale, Kedar, et al. "Disease Prediction using Machine Learning." (2019).Mr. Chala Beyene, Prof. Pooja Kamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques”, International Journal of Pure and Applied Mathematics, 2018.
9. S. Patel and H. Patel, “Survey of data mining techniques used in healthcare domain,” Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March, 2016.
10. Balasubramanian, Satyabhama, and Balaji Subramanian. "Symptom based disease prediction in medical system by using Kmeans algorithm." International Journal of Advances in Computer Science and Technology 3.
11. Dhenakaran, K. Rajalakshmi Dr SS. "Analysis of Data mining Prediction Techniques in Healthcare Management System." International Journal of Advanced Research in Computer Science and Software Engineering 5.4 (2015).
12. T. Vivekanandan and N. C. S. N. Iyengar, ‘‘Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease,’’ Comput. Biol. Med., vol. 90, pp. 125–136, Nov. 2017.
13. A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ‘‘Prediction of heart disease using machine learning,’’ in Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275–127