Disease Symptoms Prediction Application

Riya Hedaoo [1], Abhishek Chavhan[1], Roshan Bonde[1], Sudarshan Akare[1]
Mukta Wagh [2]

[1]*Student, Department of Computer Science and Engineering, Government College of Engineering, Nagpur, India.*

[2]*Professor, Department of Computer Science and Engineering, Government College of Engineering, Nagpur, India.*

Corresponding Author: roshanbonde30@gmail.com

Abstract: - People nowadays suffer from various diseases due to the environment and their lifestyle choices. As a result, predicting disease at an early stage becomes a critical responsibility. However, doctors find it challenging to make precise predictions based on symptoms. Predicting sickness accurately is the most challenging task. For the prevention and treatment of disease, accurate and timely investigation of any health-related problem is critical. In the case of a critical illness, the standard method of diagnosis may not be sufficient. The development of a Disease Symptoms Prediction Application based on machine learning (ML) algorithms for illness prediction can aid in a more accurate diagnosis than the current methods. Using Supervised machine learning techniques, we created a disease prediction system. The diagnosis system outputs the seriousness of disease that an individual may be suffering from based on the individual's symptoms, age, and gender. This application allows users to share their symptoms based on the disease chosen then it processes data to predict the disease. The system evaluates the symptoms provided by the user as input and gives the probability of the disease as an output Disease Prediction.

Key Words — *Disease Prediction, Machine Learning, Symptoms, Prediction System.*

I. INTRODUCTION

Today's individuals are more likely addicted to the web however they don't care about their personal health. During this Twenty First Century humans are a unit encircled with technology as they're the constituent of our day-to-day life cycle.[1] Chronic diseases and conditions are on the rise worldwide. An aging population and changes in societal behavior are contributing to a steady increase in these common and costly long-term health problems.[2] Also increased demands of healthcare systems due to chronic disease has become a major concern. Individuals avoid traveling to hospital for mild symptoms which can become a significant malady in future. This is responsible for increasing obesity rates and cases of diseases such as diabetes upward.

Diseases can be managed effectively with the combination of lifestyle changes, medicine and so on. The successful treatment is always traced by the right and accurate diagnosis of a patient. With the right treatment, the symptoms of disease can be reduced and improved health. The predicted results can be used to prevent and thus reduce cost for treatment and other expenses.

Medicine and healthcare 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 ugly and erratic. In this environment, when everything has gone virtual, physicians and nurses are doing everything they can to save people's lives, even if it means putting their own lives in peril. There are also some remote villages which lack medical facilities. Board-certified doctors who choose to operate online via video and phone sessions rather than in-person consultations are known as virtual doctors; however, this is not practical in an emergency. Machines are always seen to be superior to people because, in the absence of human error, they can complete jobs more quickly and consistently. A disease predictor, often known as a virtual doctor, may accurately predict the disease of any patient without the need for human

Manuscript revised April 23, 2022; accepted April 24, 2022.
Date of publication April 25, 2022.

This paper available online at www.ijprse.com

ISSN (Online): 2582-7898; SJIF: 5.59

intervention. A disease predictor may also be a benefit in situations like COVID-19 and EBOLA, since it can diagnose a human's disease without any physical touch.[3] Some virtual doctor models exist, but they do not provide the appropriate degree of accuracy since all of the criteria are not taken into account.

Regardless of what has been seen so far, it is clear that traditional tactics are falling behind somewhere. That is why a solution must be designed to close these loopholes. Machine Learning is one of the solutions for removing these loopholes. While ML vary in scale and complexity, their general structure is the same. Several rule-based techniques were drawn from machine learning to recall the development and deployment of the predictive model. Several models were initiated by using various machine learning (ML) algorithms that collected raw data and then bifurcated it according to gender, age group, and symptoms. The data-set was then processed in several ML models. According to ML models, the accuracy varied. This research is to find the best out of these best algorithms. So, for that matter some of the Machine Learning algorithms are chosen for this research. Out of those GLM, Random Forest, Naive Bayes algorithms and Principal Component Regression technique are used in this research. The output of all the models will be studied and compared and the one with best accuracy will be implemented on the final dataset. Our primary goal in this research is to create a system that can anticipate and present the key features of an illness, as well as its severity, based on the user's input of symptoms. The system may compare the symptoms to the datasets contained in the data. There are different types of methodology used in order to do disease detection and filtering. So, it is important to choose the right and suitable methodology thus it is essential to understand the application functionality.

In this study, we have used GLM, Random Forest, Naive Bayes algorithms and Principal Component Analysis techniques. First of all, we had to gather data to train our ML model. The required data for all of the diseases are collected from Kaggle datasets. After data collection, the next step in the implementation process is to pre-process the data and then train the dataset on an algorithm and after that testing it for the final prediction. So, for this purpose, RStudio was used, which provided a better coding environment and the language that is used was R. For the Diabetes dataset we have to implement Binary classification on the dataset which has two labels namely yes / positive or no / negative for prediction of diabetes. For this purpose, we have used GLM, Random Forest and Naive Bayes algorithms. The output of all three models will be studied and compared. And one with best accuracy will be implemented on the final dataset. The final output will display the predicting result on the disease.

II. Literature Review

Using machine learning algorithms, a number of studies have been conducted to predict diseases based on symptoms displayed by a person.

In the paper [4], two supervised data mining algorithms were applied on the dataset to predict the possibilities of having heart disease of a patient, and were analyzed with the classification models namely Naive Bayes Classifier and Decision tree classification. The Decision tree model has predicted the heart disease patient with an accuracy level of 91% and Naive Bayes classifier has predicted heart disease patient with an accuracy level of 87%.

In the paper [5], two different data mining classification techniques were used for the prediction of various diseases and their performance was compared in order to evaluate the best classifier. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for medical applications.

The paper [6] “Disease Prediction System” used Decision tree, Random Forest and Naive Bayes algorithms to predict a disease on the basis of systems and to enable synchronized and well-versed medical systems ensuring maximum patient satisfaction.

The paper [7] “Application of Machine Learning Predictive Models in the Chronic Disease” focused on SVM and LR algorithms and evaluated the study models associated with diagnosis of chronic disease. These models are highly applicable in classification and diagnosis of CD.

The paper [8] “Disease Prediction using Machine Learning” used KNN, Naive Bayes, Logistic Regression and Decision Tree algorithms to make a disease prediction system which can predict the disease on the basis of symptoms and implemented using grails framework.

The paper [9] “Disease Prediction using Machine Learning” used Naive Bayes, Decision Tree and Random Forest algorithms to create a disease prediction system with better accuracy and it also provides motivational thoughts and images.

In this paper [10], it is presumed that albeit most analysts are utilizing diverse classifier methods, for example, Neural system, SVM, KNN and twofold discretization with Gain Ratio Decision Tree in the conclusion of coronary illness, applying Naive Bayes and Decision tree with data pick up counts gives better outcomes in the finding of coronary illness and better exactness when contrasted with different classifiers.

The paper [11] contributes the correlative application and analysis of distinct machine learning algorithms in the R software which gives an immediate mechanism for the user to use the machine learning algorithms in R software for forecasting the cardiovascular diseases.

III. PROPOSED METHODOLOGY

3.1 Diabetes dataset:

The data was gathered by direct questionnaires from Sylhet Diabetes Hospital patients in Sylhet, Bangladesh, and was approved by a doctor. Age, gender, frequent urination, increased thirst, abrupt weight loss, weakness, increased appetite, genital thrush, visual blurring, itching, irritability, delay in healing, partial paresis, muscle stiffness, absence of hair, and obesity are among the 16 symptoms in the dataset. It has 200 negative and 320 positive remarks.

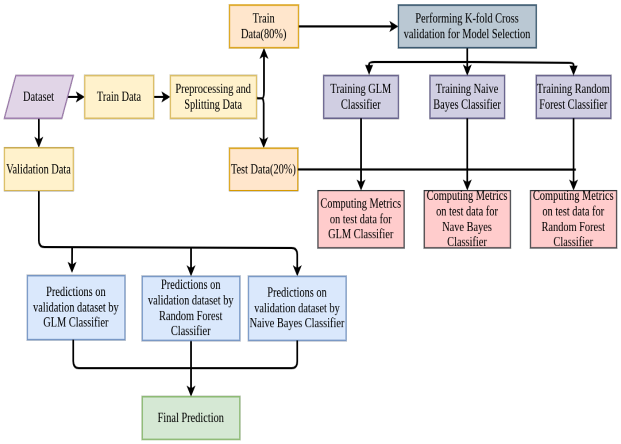

Train Data(80%)
Performing K-fold Cross validation for Model Selection
Dataset
Train Data
Preprocessing and Splitting Data
Training GLM Classifier
Training Naive Bayes Classifier
Training Random Forest Classifier
Validation Data
Test Data(20%)
Computing Metrics on test data for GLM Classifier
Computing Metrics on test data for Nave Bayes Classifier
Computing Metrics on test data for Random Forest Classifier
Predictions on validation dataset by GLM Classifier
Predictions on validation dataset by Random Forest Classifier
Predictions on validation dataset by Naive Bayes Classifier
Final Prediction

Fig.1. System Architecture

Fig. 1 shows functioning of the system. The dataset is divided into two parts: training and testing. Data preprocessing techniques were applied, and some columns were renamed to make them simple and easy to understand and access, including converting types of column values from character to categorical factors. To find outliers, results were compared for all variables. Following preprocessing, we performed binary classification to predict diabetes using the dataset, which contains two labels: Yes / Positive and No / Negative. Many algorithms are available for this purpose. We have used the GLM, Random Forest, and Naive Bayes algorithms. The output of all three models will be analyzed and compared, with the best model being used on the final dataset.

3.2 GLM Binomial Model:

Binomial Regression is a part of the Generalized Linear Model which is used to find the relationship between the independent variables and the dependent variable. It makes use of the 'logit' link function to estimate which category the observations will fall in. The GLM Binomial model was used to train the dataset, and the following results were found.

This model predicted 119 observations correctly, while 11 were inaccurate. The observed accuracy is 91.5 %, with a range estimated to be between 85% and 96%.

3.3 Random Forest Model:

Random Forest is a collection of a large number of individual decision trees. Each decision tree predicts an output, and the Random Forest model output with the most votes is chosen. Random Forest can be thought of as a decision-tree ensemble model. The Random Forest model was used to train the dataset, and the following results were found.

This model correctly predicted 127 observations, but was incorrect for only 3. The observed accuracy is 97.6%, with a range estimated to be between 93% to 99%.

3.4 Naive Bayes Model:

The Naive Bayes classifier is a group of several classifiers that use the Naive Bayes algorithm. Every variable feature is assumed to be independent of the others. We have used Naive Bayes model to train the dataset, and the research outcomes were found.

This model predicted 113 observations correctly, while 17 were inaccurate. The observed accuracy is 86.9%, whereas the range is anticipated to be between 79% and 92%.

Table.1. Algorithms Performance Analysis

Algorithm	Observed Accuracy	Accuracy Range
GLM Binomial Model	91.5%	85% - 96%
Random Forest Model	97.6%	93% - 99%
Naive Bayes Model	86.9%	79% - 92%

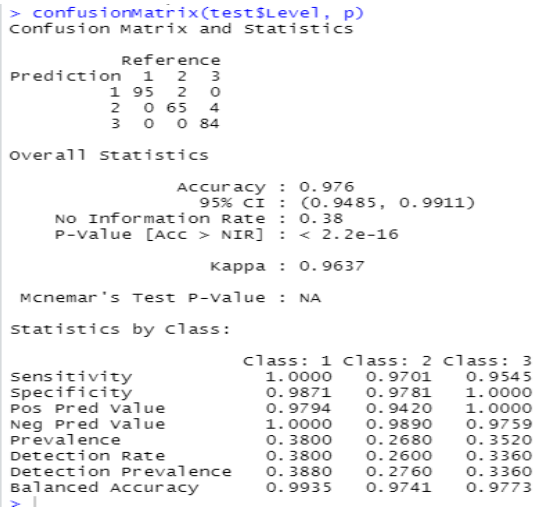
We may deduce that random forest has the highest accuracy after researching and comparing all three models. As a result, for the diabetes dataset, we will employ the Random Forest model as our final prediction model.

3.5 Cancer dataset:

The Cancer Disease dataset was used for the study, and it has a number of attributes that can be used to screen for early symptoms. There are 1000 unique rows and 25 columns in this dataset.

On the cancer dataset, the data preprocessing steps were performed. We have applied Principal Component Analysis on the cancer dataset. Principal Component Analysis (PCA) is a pattern identification technique that can be used to evaluate high-dimensional data that is difficult to comprehend just by looking at a large amount of data. For analysis of the data, we must first reduce the high data dimension to a low dimension, then plot and interpret the results.

PCA is used to present crucial data in a few simple graphs, such as a score plot and a loading plot. It is quite tough to examine big amounts of data in the field of study. To compute the relationship between the massive correlated dataset, the PCA algorithm is applied [15].



```
> confusionMatrix(test$Level, p)
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3
         1 95  2  0
         2  0 65  4
         3  0  0 84

Overall Statistics

               Accuracy : 0.976
                 95% CI : (0.9485, 0.9911)
    No Information Rate : 0.38
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9637

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            1.0000   0.9701   0.9545
Specificity            0.9871   0.9781   1.0000
Pos Pred Value         0.9794   0.9420   1.0000
Neg Pred Value         1.0000   0.9890   0.9759
Prevalence             0.3800   0.2680   0.3520
Detection Rate         0.3800   0.2600   0.3360
Detection Prevalence   0.3880   0.2760   0.3360
Balanced Accuracy      0.9935   0.9741   0.9773
> |
```

Fig.2. PCA Analysis Output

We computed principal components and utilized some of them as predictors in a linear regression model that was fitted using the typical least square method. The dataset was reduced from 25 variables to 9 PCA components. After the analysis, the observed accuracy is 97.6%, with a range estimated to be between 94% to 99%.

IV. RESULTS AND DISCUSSION

Our proposed system works as follows: the user first selects the ailment from which he or she may be suffering, and then the application directs the user to the next page, where a list of symptoms depending on the disease picked on the previous page will be displayed. Following the user's selection of appropriate symptoms, the data will be transcribed into the ML Model, which will forecast the likelihood that the user will become infected with the disease. The output will be shown to the user on the Result Page after this prediction.

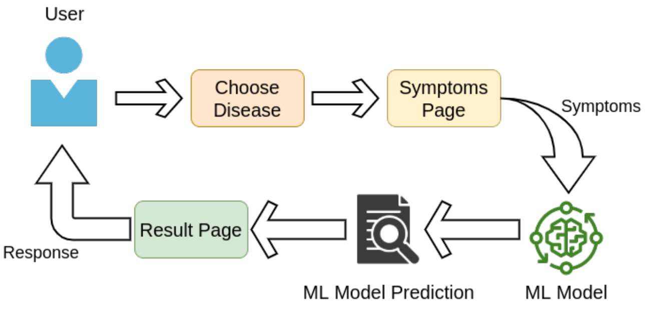


Fig.3. System Architecture of Disease Prediction Application

V. CONCLUSION

We have implemented an individual user-oriented system in the proposed system that will assist the user in forecasting the correct result for a certain condition. The developed system does not necessitate a great deal of technical knowledge to operate. Methods for predicting disease based on a patient's various symptoms were also detailed in the article. The GLM Binomial, Random Forest, and Naive Bayes algorithms were researched and applied to the diabetes dataset, and we determined that the Random Forest approach is best suited for the diabetes dataset, with an accuracy of 97.6%. The cancer dataset was also subjected to Principal Component Regression Analysis, which resulted in a 97.6% model accuracy.

Because some models were parameter-dependent, they were unable to forecast the disease and had a low accuracy rate. Once the disease is forecast, we could easily manage the medical resources required for treatment. This methodology would help to lower the cost of detecting and treating diseases while also improving recovery.

REFERENCES

[1]. Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively.

[2]. PwC, Chronic diseases and conditions are on the rise: Emerging trends: Healthcare: Industries: PwC:

[3]. Mr.Santhana Krishnan.J, Dr.Geetha.S, “Prediction of Heart Disease Using Machine Learning Algorithms”,2019 1st International Conference on Innovations in Information and Communication Technology(ICIICT).

[4]. “Multi Disease Prediction Using Data Mining Techniques” K. Gomathi , Dr. D. Shanmuga Priyaa (2017).

[5]. Khurana, Sarthak., Jain, Atishay., Kataria, Shikhar., Bhasin, Kunal., Arora, Sunny., & Gupta, Dr.Akhilesh . Das. (2019). Disease Prediction System. International Research Journal of Engineering and Technology, 6(5), 5178-5184.

[6]. Battineni, Gopi., Sagaro, Getu.Gamo. ,Chinatalapudi, Nalini.,& Amenta,Francesco. (2020). Application Of Machine Learning Predictive Models in the Chronic Disease. International of Personalised Medicine, 10(21), 1- 11.

[7]. Pingale, Kedar., Surwase, Sushant., Kulkarni, Vaibhav., Sarage, Saurabh., & Karve, Prof. Abhijeet. (2019). Disease Prediction using Machine Learning. International Research Journal of Engineering and Technology, 6(12), 2810-2813.

[8]. Chauhan Raj H., NaikDaksh N., Halpati,Rinal A., Patel,Sagarkumar J. , &PrajapatiMr. A.D. (2020). Disease Prediction and Consultation Using Machine Learning.

[9]. "Applying Machine learning methods in Diagnosing Heart disease for Diabetic Patients" G.Parthiban, S.K.Srivasta, International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 3– No.7, August 2012.

[10]. “Prediction of Cardiovascular Disease Using Machine Learning Algorithms” Dinesh Kumar G., Santhosh Kumar D. (2018).

[11]. Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.

[12]. Kaggle, Cancer Patients Dataset.

[13]. S. Sehgal, H. Singh, M. Agarwal, V. Bhasker and Shantanu, "Data analysis using principal component analysis," 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), 2014, pp. 45-48.