International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume- 8, Issue- 4, July- 2020
https://doi.org/10.21276/ijircst.2020.8.4.15
www.ijircst.org

Disease Prediction System using Support Vector Machine and Multilinear Regression

Md. Ehtisham Farooqui, Dr. Jameel Ahmad

ABSTRACT- Evolution of modern technologies like data science and machine learning has opened the path for healthcare communities and medical institutions, to detect the diseases earliest as possible and it helps to provide better patient care. Accuracy of detecting the possible diseases is reduced when we do not have complete medical data. Furthermore, certain diseases are region-based, which might cause weak disease prediction. Our body shows the symptoms when something wrong is happening within our body, sometime it may be just minor problem but sometimes we can have severe illness and if we do not take care of these symptoms at the early stage then it might be too late to cure the disease. So we are proposing a disease prediction system that can predict the possible diseases based on symptoms so it can be cured at the early stage. It saves time that is required to do the complete diagnosis of the patient and based on the suggestions provided by the system we can only get the patient diagnosed for those diseases that are required. In this paper, we are using machine learning algorithms that try to accurately predict possible diseases. The results generated by the proposed system have an accuracy of up to 87%. The system has incredible potential in anticipating the possible diseases more precisely. The main motive of this study is to help the nontechnical person and freshman doctors to make a correct opinion about the diseases.

KEYWORDS- Disease Prediction System, Machine Learning, Multilinear Regression (MLR), Support Vector Machine (SVM).

I. INTRODUCTION

It is a system that is made by the use of machine learning algorithms for guessing the possible diseases based on the patient's symptoms [1]. The growth of technology has been improving our lives so far. It provides many tools that can save millions of lives, and machine learning is one of them. Machine Learning is used to develop systems that can help us predict so many diseases based on symptoms. It can suggest the doctors, probability of the possible diseases. And diagnosis can be done based on suggestion, thus cost could be reduced.

We are living in the age of technology and nowadays humans can say that almost anything is possible with the help of technology. Today we have so many tools and methods to access information from any region of this world and Information at this age is so important that without information we would not survive. We have tools that can give us or suggest relevant information at our fingertips and the internet is one of those tools. Today billions of search queries are performed daily and sometimes there given results are relevant and sometimes they are not. In those search queries, thousands of searches are related to medical advice. People often want to know if they have any serious diseases based on their signs and symptoms. But there are no tools available to give them proper information. This research tries to give them tools so that possible disease prediction information can be provided to the end-user.

II. LITERATURE REVIEW

There have been numerous studies done related to predicting the disease using different machine learning techniques and algorithms which can be used by medical institutions. This paper reviews some of those studies done in research papers using the techniques and results used by them.

MIN CHEN et al, [1] proposed a disease prediction system in his paper where he used machine learning algorithms. In the prediction of disease, he used techniques like CNN-UDRP algorithm, CNN-MDRP algorithm, Naive Bayes, K-Nearest Neighbor, and Decision Tree. This proposed system had an accuracy of 94.8%.

Sayali Ambekar et al, [2] recommended Disease Risk Prediction and used a convolution neural network to perform the task. In this paper machine learning techniques like CNN-UDRP algorithm, Naive Bayes, and KNN algorithm are used. The system uses structured data to be trained and its accuracy reaches 82% and achieved by using Naïve Bayes.

Naganna Chetty et al, [3] developed a system that gives improved results for disease prediction and used a fuzzy approach. And used techniques like KNN classifier, Fuzzy c-means clustering, and Fuzzy KNN classifier. In this paper diabetes disease and liver, disorder prediction is done and

Manuscript received July 23, 2020

Md. Ehtisham Farooqui, Student, Department of Computer Science and Engineering, Integral University, Lucknow, India, (e-mail: mefxe01@gmail.com)

Dr. Jameel Ahmad, Assistant Professor, Department of Computer Science and Engineering, Integral University, Lucknow, India

the accuracy of Diabetes is 97.02% and Liver disorder is 96.13.

Dhiraj Dahiwade et al, [4] designed a model for prediction of the disease using approaches of machine learning and used techniques like KNN and CNN. This paper suggests disease prediction i.e. based on patient's symptoms. The accuracy of KNN is 95% and the accuracy of CNN is 98%.

Lambodar Jena et al, [5] focused on risk prediction for chronic diseases by taking advantage of distributed machine learning classifiers and used techniques like Naive Bayes and Multilayer Perceptron. This paper tries to predict Chronic-Kidney-Disease and the accuracy of Naïve Bayes and Multilayer Perceptron is 95% and 99.7% respectively.

Dhomse Kanchan B. et al, [6] studied special disease prediction utilizing principal component analysis using machine learning algorithms involving techniques like Naive Bayes classification, Decision Tree, and Support Vector Machine. The accuracy of this system is 34.89% for Diabetes and 53% for Heart disease.

Pahulpreet Singh Kohli et al, [7] suggested disease prediction by using applications and methods of machine learning and used techniques like Logistic Regression, Decision Tree, Support Vector Machine, Random Forest and Adaptive Boosting. This paper focuses on predicting Heart disease, Breast cancer, and Diabetes. The highest accuracies are obtained using Logistic Regression that is 95.71% for Breast cancer, 84.42% for Diabetes, and 87.12% for Heart disease.

Deeraj Shetty et al, [8] studied the uses of data mining for diabetes disease prediction by using Naïve Bayes and KNN algorithms. This system predicts diabetes and accuracy obtained by KNN are better than Naïve Bayes.

Rashmi G Saboji et al, [9] tried to find a scalable solution that can predict heart disease utilizing classification mining and used Random Forest Algorithm. This system presents a comparison against Naïve-Bayes classifier but Random Forest gives more accurate results with accuracy 98%.

Rati Shukla et al, [10] suggested prediction and detection for breast cancer by utilizing machine learning techniques like Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network, and KNN. In this system, the Support Vector Machine gives more accurate results than all other algorithms.

Senthilkumar Mohan et al, [11] focused on hybrid techniques in machine learning that can be used for effectively predicting heart disease and used algorithms like Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network and KNN. The accuracy of this system is 88.47%.

Anjan Nikhil Repaka et al, [12] designed and implemented a prediction model for heart disease using naive Bayesian. Any user can use this system using any smartphone device and get the prediction results. The accuracy of this system is 89.77%.

Aakash Chauhan et al, [13] proposed a disease prediction model for heart disease by utilizing evolutionary rule learning. Association Rule is used in this proposed system. This system is not very efficient because it has an accuracy of 53%.

Aditi Gavhane et al, [14] suggested prediction for heart disease that utilizes Machine Learning. Multi-Layer Perceptron model is used in this system. This system predicts heart disease based on basic symptoms like age, sex, pulse rate, etc. The accuracy of this suggested system is 91%.

III. DATASET AND MODEL DESCRIPTION

In our proposed system we are using structured datasets that can be created by collecting patient's symptoms and diagnosis from local hospitals and from open source libraries available online. We are using true datasets that gives higher accuracy.

In proposed system we utilize machine learning algorithms to predict diseases based on patient's symptoms. In this system we are predicting five diseases based on symptoms but if we feed datasets of other diseases to the system then it can also predict other diseases.

IV. PROPOSED METHOD

There are following steps involved in our proposed methodology:

i. First I will collect the datasets of symptoms and their functional problem in the body.
ii. Then I will collect the information that will associate the symptoms to possible diseases thus related disease information will be collected.
iii. Then I will get the symptoms as input from the patient and process it by Multilinear Regression.
iv. After that Multilinear Regression predicts the diseases that may be possible for those acquired symptoms.
v. Then the system will show the diagnosis in the form of max possible disease and min possible disease.

The flow chart of the methodology is given below:

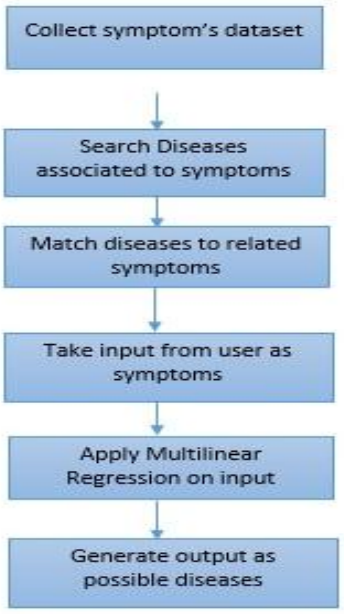

Fig 1: Flow chart of proposed method

International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume- 8, Issue- 4, July- 2020
https://doi.org/10.21276/ijircst.2020.8.4.15
www.ijircst.org

Algorithm Used:

As the name suggests, In our disease prediction system, We are using the Support Vector Machine (SVM) for classification and Multilinear Regression (MLR) for predicting the result. MLR is a form of regression algorithm where multiple independent values are involved, meaning that we try to predict a value based on two or more variables.

Simple Linear Regression, where a single Independent/Predictor(X) variable is used to model the response variable (Y). But there may be many situations in which the reply variable is affected by multiple forecaster variables; for such cases, we use the MLR algorithm.

V. THE ARCHITECTURE OF DISEASE PREDICTION SYSTEM

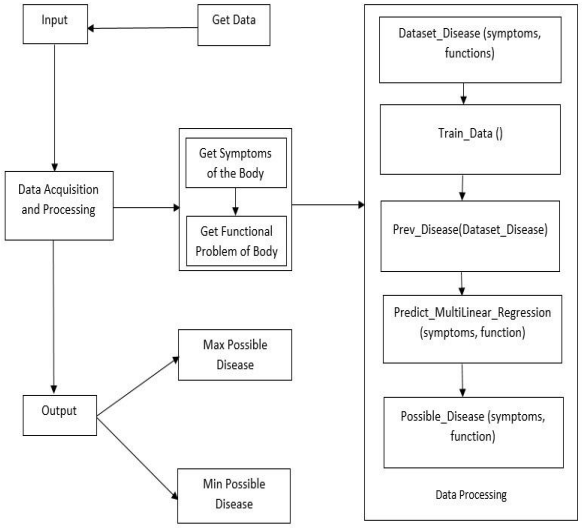

Fig 2: The architecture of Disease Prediction System

The architecture of DPS includes multiple following fields:

Input

We are taking input from the user of the disease prediction system as a symptoms list.

Get Data

In this field, the user will provide data about their symptoms.

Data Acquisition and Processing

In this field, the input is provided for processing. Data acquisition and processing perform two operations, first is the acquiring the data and then second is the processing of the data and extracting information based on that acquired data.

Get Symptoms of the Body

In this field symptoms of the body are gathered and analyzed. So that this information can be used by the algorithm to predict the possible diseases.

Get Functional Problem of Body

In this field, functional problems of the body that is associated with the symptoms are gathered. So that it is analyzed to get the possible disease.

Dataset_Disease (symptoms, functions)

In this field, we have a predefined dataset of diseases that involves symptoms and functions that are caused by the disease. This dataset is further used to match the data that has been obtained from the user and if matched properly then the system will suggest the possible diseases.

Train_Data ()

In this field training of the system is performed. Our disease prediction system is trained using the SVM (support vector machine) algorithm. Here we are using the SVM algorithm to solve a problem related to regression.

Prev_Disease (Dataset_Disease)

In this field Dataset of the diseases is provided as parameter and processing are performed based on this dataset.

Predict_MultiLinear_Regression (symptoms, function)

In this field, the prediction is performed using the MLR algorithm. In MLR, multiple independent variables are used to perform the prediction of the disease. Symptoms and their functions in the user's body are involved in the prediction.

Possible_Disease (symptoms, function)

In this field symptoms and functions are passed as a parameter and possible diseases are calculated based on these parameters.

Data Processing

This field contains the above five data processing fields and is the main part of our disease prediction system. It has all the necessary fields for processing the data.

Output

After Data Acquisition and Processing, possible diseases are generated as output.

Max Possible Disease

This field contains the maximum possible disease as output.

Min Possible Disease

This field contains the minimum possible diseases as output.

VI. ALGORITHM FOR DISEASE PREDICTION SYSTEM

The algorithm that we are using in our proposed system is given below and by using this algorithm we are getting the accuracy of up to 87%.

i. Take input of symptoms in p[ ] and their function in t[ ].
ii. Declaration:
    S[n][m] - m set of symptoms of n disease
    F[n][m] - m set of functions of n disease
iii. $\left|\sum_{i \in I} D(i)\right| \leq \sigma$, and it has a minimum cardinality.
iv. Set Ssvm[ ] = new set of possible disease symptoms
v. Set Fsvm[ ] = new set of possible disease functions
vi. For each x, y in Ssvm[ ] and Fsvm[ ]
vii. If p[ ] in Ssvm[ ] and t[ ] in Fsvm[ ]

viii. Pdisease[ ] = Ssvm[ ], Fsvm[ ], Priority++
ix. endIf
x. endFor
xi. Possible disease Pdisease[0]

Explanation of the above-written algorithm:

- Input is taken from the user in the form of symptom and stored in p[ ] and function is stored in t[ ].
- In the second step, the declaration is done and S[n][m] stores m set of symptoms of n disease and F[n][m] stores m set of functions of n disease.
- $\left|\sum_{i \in I} D(i) \leq \sigma\right.$, it has a minimum cardinality.
- Set Ssvm[ ] is a new set of possible disease symptoms
- Set Fsvm[ ] = new set of possible disease functions
- In this step, we are using a for loop to check each value of Ssvm[ ] and Fsvm[ ]
- And if given input lies within Ssvm[ ] and Fsvm[ ].
- Then we are storing that disease in Pdisease[ ] and increasing the priority of that disease.
- If statement end.
- For loop end.
- The possible disease is given as output Pdisease[0].

VII. RESULT ANALYSIS

Result analysis in our proposed system is an essential part of this research paper. By the analysis of results we can compare that how much better this proposed system is performing. In result analysis we will see accuracy of different diseases that are predicted using our proposed system. We have taken datasets of 100 cases for result analysis.

Disease based accuracy analysis for 100 cases:

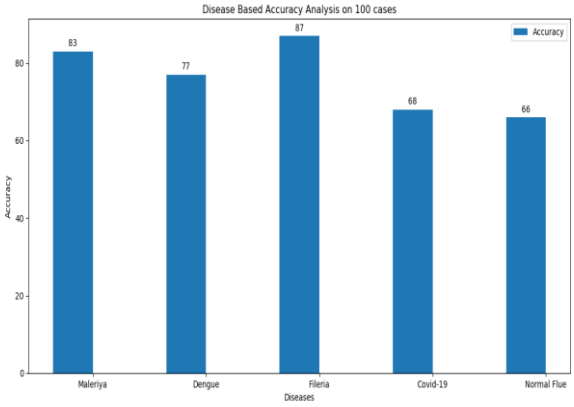

Fig 3: Disease Based Accuracy analysis on 100 cases

Above diagram shows the accuracy of 5 diseases that are malaria, dengue, filaria, covid-19 and normal flu.

Disease based accuracy analysis for 100 cases using SVM and CNN:

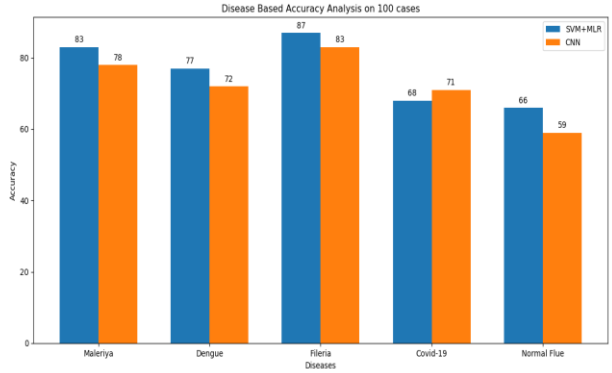

Fig 4: Disease Based Accuracy Analysis on 100 cases comparison

In the above chart we can see that five diseases are given and for these 5 diseases there accuracies are also given. These five diseases are processed using two different algorithms for each consecutive bars. The blue bar shows accuracy for the diseases processed using SVM. The Orange bar shows the accuracy of diseases processed using CNN.

Response Time Analysis:

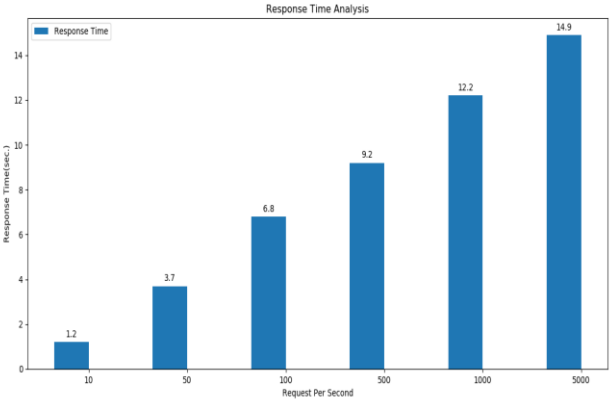

Fig 5: Response Time Analysis

In the above chart x axis shows request per second and y axis shows response time in seconds. Result analysis for response time is very important to show that our system can handle multiple requests at a time.

Comparative analysis between algorithms for our disease prediction system:

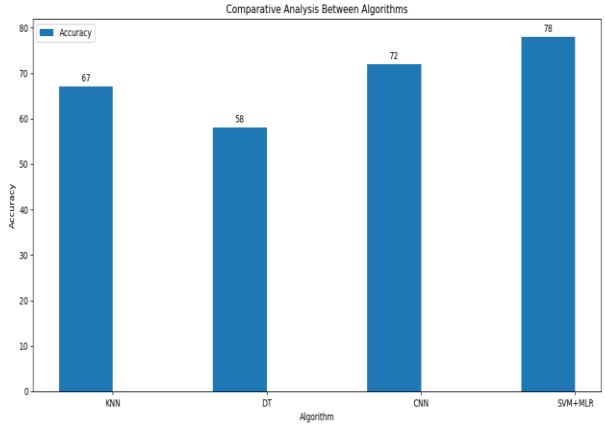

Fig 6: Comparative Analysis between Algorithms

Combination of SVM and MLR is performing better that other algorithms.

VIII. CONCLUSION AND FUTURE WORK

In our research, we have used a support vector machine and multilinear regression algorithm to predict diseases. And we have also tested multiple algorithms like the k-nearest neighbor, convolution neural network, decision tree, etc. Despite testing these algorithms I have found that the support vector machine and multilinear regression combination gives higher accuracy than other algorithms.

The purpose of this research was to provide medical diagnosis information based on symptoms to normal people, fresher doctors, medical students, and anyone who wants to know about a set of symptoms and associated diseases.

In this research, we have found that possible disease prediction can go up to 87% for some diseases and minimum 68% for some diseases but these results are obtained using the minimum amount of data set but if we can feed the system humongous amount of data set then this disease prediction system can give accuracy up to 95%. Obtaining a humongous amount of data set related to diseases and their symptoms is very time consuming and it cannot be done within one or two years it requires multiple years to collect those data sets and train the system using those data searches. This system can be used by Ph.D. scholars to do further research.

With the help of a disease prediction system, it was possible to diagnose people based on symptoms. Disease prediction system provides only possible outcomes it does not guarantee that it will predict the disease correctly. But it has significantly higher accuracy for predicting possible diseases. In our research, we have analyzed the accuracy of this system for 5 different diseases and our accuracy can go up to 87%.

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] **S**ayali Ambekar, Rashmi Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network" IEEE, 978-1-5386-5257-2/18, 2018.

[3] Naganna Chetty, Kunwar Singh Vaisla and Nagamma Patil, "An Improved Method for Disease Prediction using Fuzzy Approach" IEEE, DOI 10.1109/ICACCE.2015.67, pp. 569-572, 2015.

[4] Dhiraj Dahiwade, Gajanan Patle and Ektaa Meshram, "Designing Disease Prediction Model Using Machine Learning Approach" IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4, pp. 1211-1215, 2019.

[5] Lambodar Jena and Ramakrushna Swain, "Chronic Disease Risk Prediction using Distributed Machine Learning Classifiers" IEEE, 978-1-5386-2924-6/17, pp. 170-173, 2017.

[6] Dhomse Kanchan B. and Mahale Kishor M., "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis" IEEE, 978-1-5090-0467-6/16, pp. 5-10, 2016.

[7] Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Disease Prediction" IEEE, 978-1-5386-6947-1/18, pp. 1-4, 2018.

[8] Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil, " Diabetes Disease Prediction Using Data Mining" IEEE, 978-1-5090-3294-5/17, 2017.

[9] Rashmi G Saboji and Prem Kumar Ramesh,"A Scalable Solution for Heart Disease Prediction using Classification Mining Technique" IEEE, 978-1-5386-1887-5/17, pp. 1780-1785, 2017.

[10] Rati Shukla, Vikash Yadav, Parashu Ram Pal and Pankaj Pathak, "Machine Learning Techniques for Detecting and Predicting Breast Cancer" IJITEE, ISSN: 2278-3075, Volume-8, pp. 2658-2662, 2019.

[11] Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access, DOI 10.1109/ACCESS.2019.2923707, pp. 81542-81554, 2019.

[12] Anjan Nikhil Repaka, Sai Deepak Ravikanti and Ramya G Franklin,"Design And Implementing Heart Disease Prediction Using Naives Bayesian" IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8, pp. 292-297, 2019.

[13] Aakash Chauhan, Purushottam Sharma, Vikas Deep and Aditya Jain, "Heart Disease Prediction using Evolutionary Rule Learning" CICT 2018.

[14] Aditi Gavhane, Gouthami Kokkula, Isha Pandya and Kailas Devadkar, "Prediction of Heart Disease Using Machine Learning" IEEE Xplore ISBN: 978-1-5386-0965-1, pp. 1275-1278, 2018.

[15] Ankita Dewan and Meghna Sharma, "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification" IEEE, 978-9-3805-4416-8/15, pp. 704-706, 2015.

ABOUT THE AUTHORS

Md. Ehtisham Farooqui has done B.Tech in Computer Science from Institute of Technology and Management, GIDA, Gorakhpur, India and currently perusing M.Tech from Integral University, Lucknow, India

Dr. Jameel Ahmad is Assistant Professor, Department of Computer Science and Engineering in Integral University and has more than 18 years of teaching experience.