IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 5, May 2022

DOI: 10.17148/IARJSET.2022.9594

Disease Prediction System and Covid Prediction Probability Using ML

Gaurav Birdi[1], Sachin Sharma[2], Md. Omer[3], Prakhar Katiyar[4], Prof. Ms. Upasna Joshi[5]

[1-5]Department of Computer Science and Engineering, Delhi Technical Campus, Greater Noida, India

Abstract: The prediction of disease precisely at an early stage is crucial in order to provide an efficient treatment. The conventional way of diagnosing disease can be inefficient in such circumstances. With the advancement of Machine Learning, a system can be designed for increasing the accuracy of the disease prediction. This goal can be achieved by using the various Machine Learning Algorithms available. The available dataset provides us with the information regarding the symptoms of 50 diseases. Also, there is another dataset which provides us with symptoms related to Covid-19 virus. In general disease prediction the average accuracy of all the algorithms is 94.6% and in the Covid-19 dataset the accuracy is 92.5%. This diagnosis system can act as doctor's assistant or a pre-diagnosis agent for the patients. Lives can be saved with the possibility of an early diagnosis of a life-threatening disease.

Keywords Disease prediction, Covid-19, Machine Learning Algorithms

1 INTRODUCTION

Now a days, people face various diseases due to the environmental condition and their living habits. So, the prediction of disease at earlier stage becomes important task. But the accurate prediction on the basis of symptoms becomes too difficult for doctor. The correct prediction of disease is the most challenging task. To overcome this problem data mining plays an important role to predict the disease. Medical science has large amount of data growth per year. Due to increase amount of data growth in medical and healthcare field the accurate analysis on medical data which has been benefits from early patient care. With the help of disease data, data mining finds hidden pattern information in the huge amount of medical data. We proposed general disease prediction based on symptoms of the patient. Disease Prediction helps patients to identify the risk of disease or health conditions

We proposed general disease prediction based on symptoms of the patient. Disease Prediction helps patients to identify the risk of disease or health conditions. In present scenario, due to the pandemic people feel unsafe to visit the hospital and if patient is not serious and just wants to identify the type of ailment, disease prediction plays a key role. Therefore, a need for a system arises with the help of which users/patients can diagnose the disease on the basis of symptoms that they are facing, which would give them an awareness of the disease they are suffering from, medication needed, specific doctor to be consulted and further avoiding potential hospital admissions.

Virtual Doctors are available which are certified and opted to provide patients with diagnosis using online portals through certain applications. But this is rather difficult in emergency cases as you are supposed to book an appointment first. In such cases a computer system that is capable of diagnosing a patient while a doctor is not available can make a difference between life or death. Also, in the situations like pandemics such as COVID – 19, a virus that can spread rapidly and easily through human contact, a machine learning based system which is immune towards such diseases is a liable choice for the diagnosis. Some models exist which predict diseases related to a certain organ which can prove to be inefficient at times. So, having a system that can predict diseases through a large spectrum and with a certain accuracy that is acceptable in this case can be considered a better choice. Several machine leaning techniques can be used for the prediction of diseases such as: Logistic Regression, Decision Tree, Random Forest, Gaussian Naïve Bayes, Support Vector Classification and KNN can be used to create an accurate model for the prediction of a vast spectrum of diseases.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 5, May 2022

DOI: 10.17148/IARJSET.2022.9594

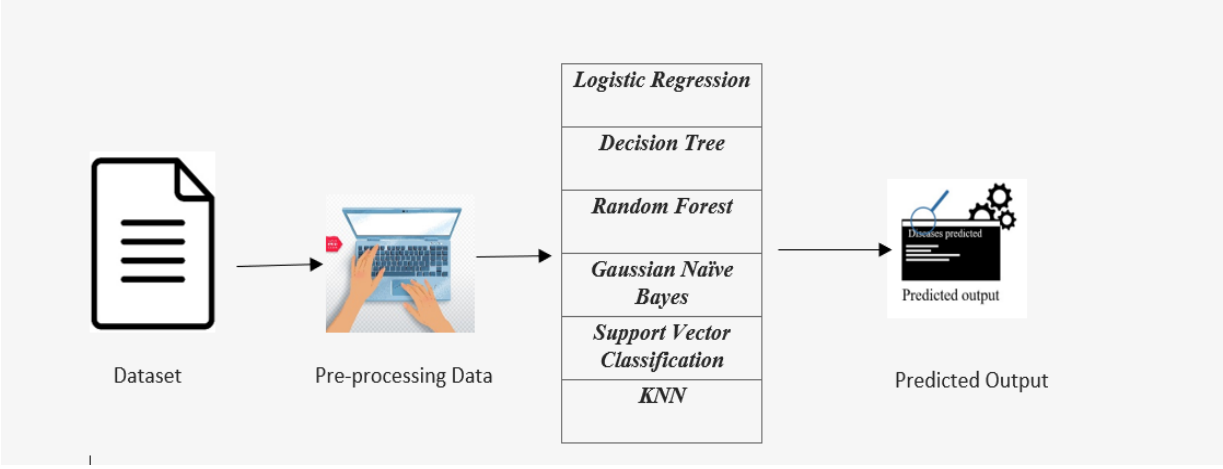

Fig. 1 Working system flow diagram during training. There are columns containing diseases, their symptoms, precautions to be taken, and their weights and after cleaning the data it was processed into different machine learning algorithm such as Logistic Regression, Decision Tree, Random Forest, Gaussian Naïve Bayes, support vector classification and KNN for the disease prediction. The final result is the disease predicted on the basis of symptoms that were trained with the help of the dataset.

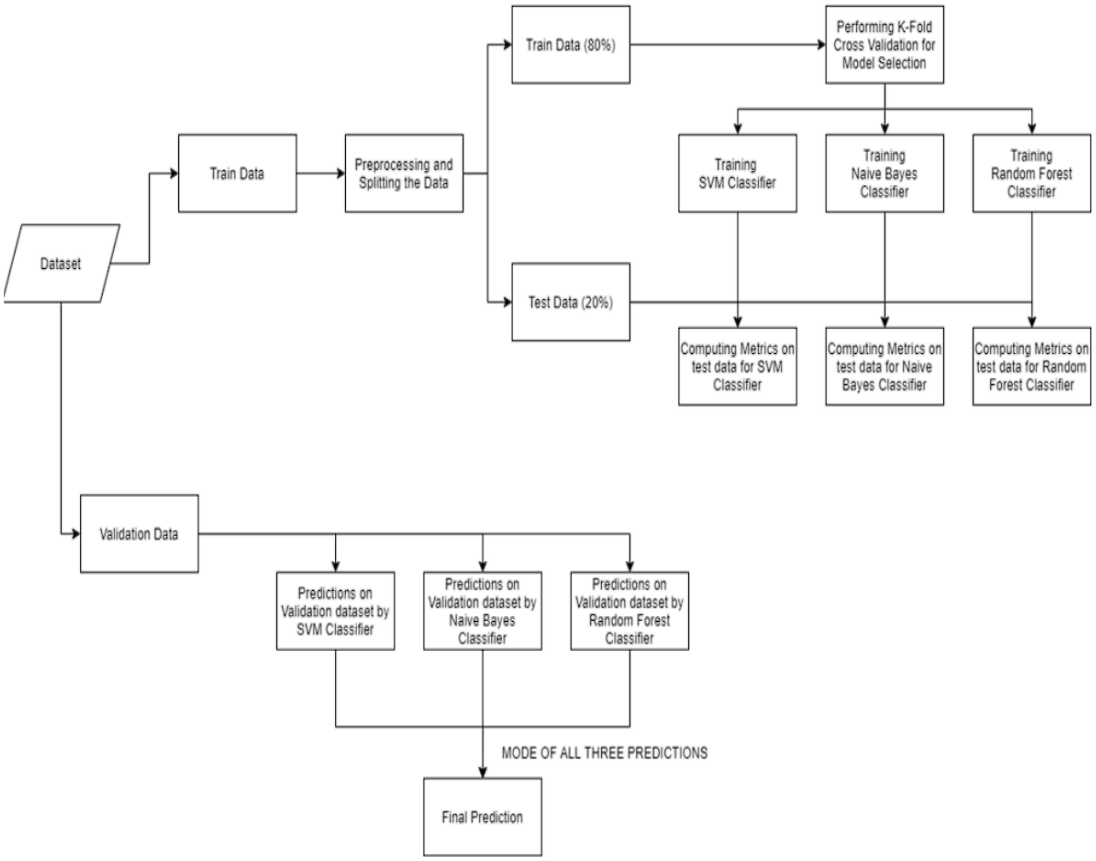

Fig. 2 here the dataset is divided into two parts for training and validation, the training data is preprocessed and its splits in train data (80%) and test data (20%). The train data performs k-fold cross validation and three models are train i.e. Random Forest, Gaussian Naïve Bayes, SVM classifier. Concurrently metrics are computed on test data for svm, naïve bayes, random forest classifier simultaneously prediction on validation data are performed by svm, random forest and naïve bayes classifier and the mode of all the three predictions is taken and a final prediction is computed.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 5, May 2022

DOI: 10.17148/IARJSET.2022.9594

Algorithm	Accuracy
Logistic Regression	95%
Decision Tree	93%
Random Forest	95%
Gaussian Naïve Bayes	95%
Support Vector Classification	95%

Fig. 3 The concurrent Accuracy of all the algorithms are mentioned as observed.

2. LITERATURE SURVEY

Harish Rajora, N. Punn, S. K. Sonbhadra, Sonali Agarwal. [1] In their paper implemented a web app. In their work, the prediction is done using four different machine learning algorithms (Random Forest, Naïve Bayes, KNN and Ensemble). The source of data is not mentioned. The accuracy achieved in this paper is between 84 and 93.5 for different algorithms. Mohan Kumar K N, S.Sampath, Mohammed Imran [2] In their paper discuss the need for an affordable disease prediction system which can identify diseases in the early stage. It also lists the techniques currently being used to predict disease. Logistic regression, Support Vector Machine (SVM), Decision tree and Clustering techniques dominate with 27.5%, 25%, 22.5% and 20%. Logistic regression dominates the list because of the nature of medical data which is binary in most of the cases.

Md Ekramul Hossain, Arif Khan, Mohammad Ali Moni, Shahadat Uddin [3] In their paper discuss the different methods already used for specific disease prediction and their accuracies. The term 'Electronic Health Data' refers to the digitised health data which contains information of diseases, diagnostics and treatments of patients. All the papers with the keywords 'Electronic Health data' and 'Disease prediction' were compared and the listed in this comprehensive literature review.
Ashish Kumar, Priya Ghansela , Purnima Soni , Chirag Goswami , Parasmani Sharma [4] contains the prediction of diseases taken from several sources like hospitals, discharge slips of patients and from UCI repository. Then it applies supervised machine learning algorithms such as Decision tree, Random Forest, SVM (Support Vector Machine) and Naive-Bayes to train the model. The accuracies range from 54 in SVM to 95 in Random Forest. The exact source of the data is not given in this paper.

Min Chen, Yixue Hao , Kai Hwang , Lin Wang , Lu Wang [5] suggest the use of a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP). The data is collected from real-life hospital data from central China in 2013-2015. The prediction accuracy of the proposed algorithm reaches 94.8%.

Ch Aishwarya, K Suvarchala, B Aravind, G Shashank [6] focuses on predicting a a certain disease before it occurs. As We know that prevention is better than cure, so much so it is very important to diagnose a specific disease and follow the required guidelines before its validity iyanda. So, we came up with the idea that predicting the disease before it happens. There is already a system in place for how to do it disease prediction, but focuses on data sets only found in local health care communities (structured data). The proposed system in particular focuses on big data that is widely used today. It uses CNN with many models random data combination algorithm and systematic data. This is an algorithmic model it basically consists of three layers, namely input layer, hidden layer (can be plural) and the output layer..
Marouane Fethi Ferjani [7] explores the proposed idea that ML-monitored algorithms can improve health care with accurate and faster diagnosis. In this study, we investigate studies using more than one monitored ML model for each diagnostic problem. This approach provides more insight and accuracy because the performance testing of a single algorithm in different research settings creates biases that produce vague results. Analysis of ML models will be

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 5, May 2022

DOI: 10.17148/IARJSET.2022.9594

performed on a few diseases of the heart, kidneys, breast, and brain. To diagnose the disease, several methods will be tested such as KNN, NB, DT, CNN, SVM, and LR.

Dhiraj Dahiwade, Prof Gajanan Patle, Prof Ektaa Meshraam [8] made a prediction of three diseases such as diabetes, mental illness and heart disease. Disease prognosis is done with systematic data. Predictors of heart disease, diabetes and brain disorders are made using a different machine learning algorithm such as naïve bayes, decision tree and KNN algorithm. The result of the Decision Tree algorithm is better than the Naïve bayes and the KNN algorithm. Also, they predict whether the patient is at high risk for cerebral infarction or low risk of cerebral infarction. To predict the risk of brain damage, they used CNN multimodel disease risk prediction in text data. Accurate comparisons occur between CNN's unimodel risk predictors based on CNN based multimodel disease risk algorithm. Accuracy of disease forecasts reaches 94.8% faster than CNN-based disease risk prediction algorithm.

Rinkal Keniya, Aman Khakharia, Vruddhi Shah, Vrushabh Gada, Ruchi Manjalkar, Tirth Thaker, Mahesh Warang, Ninad Mehendale [9] Proposed disease reporting system. A doctor may not always be available when needed. However, in the current context, one can necessarily use this predictive system at any time. Individual characteristics and age and gender can be assigned to the ML model for further processing. After initial data processing, the ML model uses the current input, trains and evaluates the algorithm that leads to the predicted disease.

Dong Jin Park, Min Woo Park, Homin Lee, Young-Jin Kim, Yeongsic Kim, Young Hoon Park [10] aimed to build a new optimized ensemble model by blending a DNN (deep neural network) model with two ML models for disease prediction using laboratory test results. 86 attributes (laboratory tests) were selected from datasets based on value counts, clinical importance-related features, and missing values. The optimized ensemble model achieved an F1- score of 81% and prediction accuracy of 92% for the five most common diseases. The deep learning and ML models showed differences in predictive power and disease classification patterns.

3. METHODOLOGY

With the help of the open-source dataset from Kaggle and Data.gov.il, all the symptoms for the respective diseases are derived. We derived around 50 diseases with the support of 132 symptoms from the dataset. The symptoms were processed into machine learning algorithms

3.1 Logistic Regression

The logistic regression algorithm will provide us with binary outcome variable. It simply was used to determine the possibility of a person suffering from Covid.

The function for logistic regression is $Sig(x)=1/1+e^{-x}$.

Similarly in general purpose disease prediction system logistic regression is used to determine the possibility of an individual suffering from an disease based on symptoms in the form of binary input variables provided by the users.

3.2 Decision Tree

Decision tree algorithm is an essential part of supervised learning algorithms as it is used for regression in classification. A root node is there in a decision tree after which it gets divided in the dominant input feature and divides again. The iteration is repeated until all the inputs are filled in the node.

Maximum splits are 4 from each node of a coarse tree similarly in a medium tree such splits are around 20 and in a fine tree it is around 100.

3.3 Random Forest

Rather than being dependent on one decision tree random forest takes predictions from all the tree and based on the majority of prediction, final output is generated.

Random forest consists of number of decision tree and it takes the average of all the tree outputs to improve the accuracy. So to improve the accuracy of the final result that was generated with the help of decision trees random forest is used.

3.4 Support vector classification(svc)

Svm is one of the most used and popular supervised learning algorithms used for regression problems.

Svm creates the marginal hyperplane by dividing the datasets into number of classes with the aim of deriving maximum placements in hyperplane and obtain possible outcome from the hyperplane class.

The symptoms are placed on the hyperplane of different classed that belong to different diseases. Svm determines the particular class of disease which has maximum number of symptom present on the hyperplane and then it selects the particular class i.e. Disease and the result is derived.

3.5 K nearest neighbors (KNN)

KNN algorithm is an essential part of supervised machine learning algorithm. It calculates the distance between a new data point and all other training data type. The distance is of two types: Euclidian and Manhattan, After selecting K-nearest data point where k is an integer, the class with maximum k data points is assigned.

3.6 Gaussian Naïve Bayes

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 5, May 2022

DOI: 10.17148/IARJSET.2022.9594

Gaussian distribution is also known as normal distribution and it is extension of naïve bayes also it is easy to work with because only the mean and standard deviation is derived from training data.
Our dataset consisted of symptoms and weights.
Using these as inputs we calculated the mean and standard deviation for each disease class So, this algorithm was preferable. The accuracy of this model is quite low

4. RESULT

A number of machine learning models were used to proceed with the prediction of disease for the give input dataset. We used 6 different models for the prediction. Out of the 6 models we managed to get 90% or above accuracy for 3 models. As depicted in Figure 3, among all the models, we gained the lowest accuracy for the Decision Tree model with accuracy up to 93%. The modded accuracy is 95% for the remaining models.
Also, for the covid detection model, the accuracy of the system is upto 95%.
Logistic Regression is the most preferrable algorithm for the covid detection system and for the general disease prediction system a number of algorithms are used i.e., KNN, SVM, Logistic Regression, Decision Tree, Random Forest.

5. CONCLUSION AND FUTURE WORK

So finally, we conclude by saying that this project will help in improving future of the public health care system and help patients in getting fast and appropriate medical opinion as it provides prediction for General Disease and Covid using symptoms entered by the user at the comfort of home. The system uses Logistic Regression for Covid Prediction as Logistic Regression is an easily interpretable classification technique that gives the probability of an event occurring, not just the predicted classification, and multiple machine learning algorithms for General Disease Prediction. Using Algorithms, the maximum accuracy achieved is 95%. Timely prediction and diagnosis of an ailment can evade a general disease turning into a fatal disease.

Future Work
The Future Vision of the Project encompasses collecting information and data of latest diseases. we also plan to include as part of the project recommendation of the medical experts based on the predicted disease and develop a mobile application for the current based web application.

6. ACKNOWLEDGEMENT

We express our sincere gratitude to **Dr. Seema Verma (HOD, CSE)** and **Ms. Upasna Joshi (A.P, CSE),** Delhi Technical Campus, with their valuable guidance and timely suggestions throughout my writing career, otherwise this work would not have been possible. We would also like to extend my deepest greetings to all the other members of the Department of Computer Science, who have given their great effort and guidance at the right times otherwise it would have been very difficult for me to complete this task. Finally, we would also like to thank my friends for their advice and point out my mistakes.

7. REFERENCES

1) Rajora, H., Singh Punn, N., Sonbhadra, S. K., & Agarwal, S. (2021). Web based disease prediction and recommender system. arXiv e-prints, arXiv-2106.
2) Mohan Kumar, K. N., Sampath, S., & Imran, M. (2019). An overview on disease prediction for preventive care of health deterioration. IJEAT, 8(5S), 255-261.
3) Hossain, M. E., Khan, A., Moni, M. A., & Uddin, S. (2019). Use of electronic health data for disease prediction: A comprehensive literature review. IEEE/ACM transactions on computational biology and bioinformatics, 18(2), 745-758.
4) Ashish Kumar, Priya Ghansela, Purnima Soni , Chirag Goswami , Parasmani Sharma. (2020). Prediction of diseases using supervised learning. International journal of creative research thoughts.
5) Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, 8869-8879.
6) Aishwarya, C., Suvarchala, K., Aravind, B., Shashank, G., Anand, M., & Krishna Rao, N. V. (2020). Prediction of disease using machine learning and deep learning. In Energy Systems, Drives and Automations (pp. 69-79). Springer, Singapore.
7) Ferjani, M. F. Disease Prediction Using Machine Learning.

DOI: 10.17148/IARJSET.2022.9594

8) Dahiwade, D., Patle, G., & Meshram, E. (2019, March). Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215). IEEE.

9) Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., ... & Mehendale, N. (2020). Disease prediction from various symptoms using machine learning. Available at SSRN 3661426.

10) Park, D. J., Park, M. W., Lee, H., Kim, Y. J., Kim, Y., & Park, Y. H. (2021). Development of machine learning model for diagnostic disease prediction based on laboratory tests. Scientific reports, 11(1), 1-11.

11) Thakral, B., Saluja, K., Bajpai, M., Sharma, R. R., & Marwaha, N. (2005). Importance of weak ABO subgroups. Laboratory Medicine, 36(1), 32-34.

12) Kaur, R., & Jain, A. (2012). Rare blood donor program in the country: Right time to start. Asian Journal of Transfusion Science, 6(1), 1.