Disease Prediction Using Various Classifier Algorithms

Prof.Harshavardhan Dodmani
Department of Computer Science and Engineering
S J C Institute of Technology
Chickaballapura, B.B. Road
India
562101 sjcit.ac.in

Nisarga B
Department of Computer Science and Engineering
S J C Institute of Technology
Chickaballapura, B.B. Road
India
562101 sjcit.ac.in

Pruthvi V
Department of Computer Science and Engineering
S J C Institue of Technology
Chickaballapura, B.B. Road
India
562101 sjcit.ac.in

Shashank H P
Department of Computer Science And Engineering
S J C Institute of Technology
Chickaballapura, B.B. Road
India
562101 sjcit.ac.in

Vishal R
Department of Computer Science And Engineering
S J C Institute of Technology
Chickaballapura, B.B. Road
India
562101 sjcit.ac.in

Abstract— Machine Learning calculations are PC programs that attempt to anticipate disease type in view of the past information. The possible objective of AI calculations in malignant growth finding is to have a prepared calculation that gives the quality articulation levels from disease patient, can precisely anticipate what type and seriousness of disease they have, helping the specialist in treating it. The current innovation analyzes four different ML algorithms are Decision Tree, Random Forest, K nearest Neighbor and Naïve Bayes. A sample data consists of 42 diseases.

Keywords—Machine Learning, are Decision Tree, Random Forest, K nearest Neighbor and Naïve Bayes

I INTRODUCTION

The identification of patterns and detection of correlations and relationships within huge databases using different machine learning approaches has changed healthcare organizations. Providing data for future actions, it is an important tool in the medical sector. Using this technology, it is possible to delve deep into enormous volumes of data by combining a variety of analytical methodologies and modern algorithms. Systematically importing, organizing, and analyzing patient data is a common practice in healthcare. As a platform for a deeper understanding of the mechanisms in each element of the medical domain, it may be used to pinpoint inherent inefficiencies and provide better services. This may result in better diagnosis and medicine, and more importantly a successful treatment.

The use of machine learning approaches is making significant progress in improving the diagnostic process' efficiency by intelligently converting accessible information into value. In the area of diagnostic abilities, machine learning has been explored in numerous studies. Machine learning algorithms, according to a study, can identify with 91.1% accuracy, which exceeds the best doctor's ability to diagnose at 79.97%. In order to reliably diagnose, predict, prevent, and treat illnesses, machine learning techniques are employed explicitly to extract features.

Currently, humans can achieve almost anything with technology and we are living in the age of information. Information is so important today that we could not survive without it, thanks to the many tools and methods we have to access information from anywhere on earth. With the internet, we have a variety of tools in our hands that can help us discover relevant information. Every day tens of billions of searches are conducted, and sometimes the results provided are relevant, and sometimes they are not. Medical advice is among the many searches conducted. The signs and symptoms of serious diseases often make people wonder if they have them. The problem is that people do not have access to an adequate means of information. Hence, we propose a tool that can be used to provide end-users with possible disease prediction information using machine learning and this tool are Decision Tree, Random Forest, K nearest Neighbor and Naïve Bayes.
The machine learning algorithms, it uses a patient's symptoms to determine if a disease is present. The use of Machine Learning helps us create systems that predict many diseases based on symptoms. The patients are given suggestions about which diseases are likely to occur. The cost could also be reduced since diagnosis can be done based on suggestions.

II LITERATURE SURVEY

Using machine learning, statistics, and database systems, data mining aims to extract and discover patterns in large data sets. The efficient analysis of medical databases has aided in the early detection of illnesses. A classifier system based on Machine Learning algorithms has aided healthcare professionals in predicting and diagnosing health-related issues at early stages. A sample of data consisting of 3218 patients records who are diagnosed with 45 diseases, 97 out of 142 are independent variable diseases that are closely linked were chosen. The 3 machine learning algorithms used in the base paper were Random Forest Classifier, Decision Tree Classifier and Naive Bayes classifier. Based on the history of Machine Learning and its applications in medical diagnosis, it can be seen that they've developed systems, algorithms, and methodologies to allow more sophisticate analysis of data. The enormous amount of data generated and stored by recent technologies will make data analysis even more important in the future. Medical practitioners today can discover interesting relationships in their data using machine learning algorithms that are currently available. who are solving current and upcoming problems. Combining classifiers was found to be an effective method for improving a diagnostic system's reliability and understandability by physicians. Although clearly demonstrated technical possibilities, machine learning in medical diagnosis has yet to be accepted to the extent it can be. Despite this, one cannot expect that this disparity between practical applications and technical possibilities will endure for very long .

New multimodal disease risk prediction algorithms that use convolutional neural networks are being developed by hospitals. There is currently no existing work in the area of medical big data analytics that focuses on both types of data. This paper gives a clear picture that Machine learning algorithms are used to predict the diseases the patients are suffering from. Naïve Bayes, Random Forest, Decision Tree and KNN algorithms produce the results accurately when compared with other algorithms. Random Forest and Naïve Bayes algorithms are able to predict accurately even with less information available.

III PROPOSED METHODS FOR DISEASE PREDICTION

Through available machine learning algorithms, the disease prediction system has been implemented. The algorithms used are Decision Tree, Random Forest, K nearest Neighbor and Naïve Bayes. A detailed explanation of the algorithms is given below.

A. Decision Tree

An effective decision-tree approach involves identifying the learned capacity in the form of a choice tree for approximating the discrete-evaluated target capacity. As a revision of learned trees, you can reframe them as sets of if-then events to consider human coherence.

.

The best algorithm that is used under decision tree is ID3. This algorithm initially finds out the root node. Once root node is found subsequent nodes are found and tree structure is built. To measure the worth of the attribute information gain is used. We calculate the entropy. Suppose there are X sets of data then entropy is calculated by as follow*:*

Entropy (X)=-p_+ $\log_2 p_+$ - p_- $\log_2 p_-$
In the next step information gain is calculated.

Let us understand the algorithm clearly by taking an example. Consider the table

Blurred and distorted vision	Restlessness	Lethargy	Irregular periods	disease
0	0	1	1	Hypothyroidism
1	1	0	1	Hyperthyroidism
0	1	0	1	Hyperthyroidism
0	0	1	1	Hypothyroidism
1	0	1	1	Hypothyroidism
0	0	1	1	Hypothyroidism
1	1	0	1	Hyperthyroidism
0	1	0	1	Hyperthyroidism
1	1	0	1	Hyperthyroidism

Consider the table.

Step 1:
Calculate E(X)
Using the above formula

E(X)=-6/10 $\log_2$ (6/10)-4/10 $\log_2$(4/10)
=0.9710

Step 2:

Calculate the information gain

E(X, irregular periods)=

The total number of positives and negatives are as shown in below

1	0
10	0

Since we have only positive information gain is 0.

E(X, lathargy)=

The total number of positives and negatives are as shown in below

1	0
6	4

E(X, latency)= -4/10 $\log_2$(4/10)- 6/10 $\log_2$ (6/10)

= 0.9710

E(X, Restlessness)=

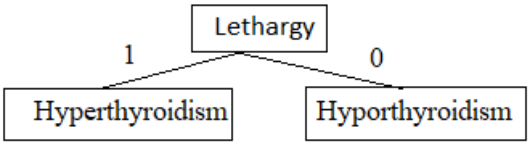
The total number of positives and negatives are as shown in below:

1	0
6	4

E(X, latency)= -4/10 $\log_2$(4/10)- 6/10 $\log_2$ (6/10)

= 0.9710

Now we try to construct the tree



B. Random Forest

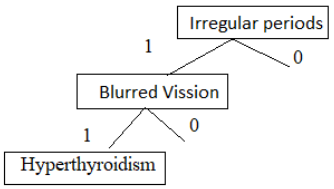
Random forest is used for classification and regression problems. It constructs decision trees on various examples and takes their greater part vote in favor of characterization and normal if there should be an occurrence of regression.

This algorithm takes the set of data. It then calculates the result of the selected set. Similarly, many sets are reviewed, and the result is found. All the results are analyzed and the major voting or the average result has been displayed.

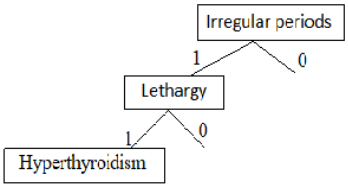
Let us try to understand random forest algorithm using below example.

Let us try to construct 3 different trees

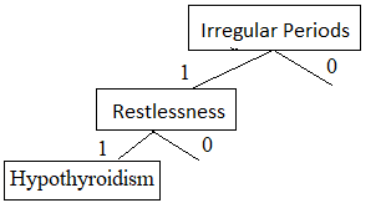
Tree -1



Tree -2



Tree -3



Now when the user gives lethargy, blurred vision and irregular periods as input, the disease is predicted as Hyperthyroidism.

C. KNN algorithm

In terms of data mining, KNN is among the most straightforward methods. The training examples need to be stored in the memory during the runtime of Memory-Based Classification.

In the case of continuous attributes, the Euclidean distance is used to calculate the difference between the attributes. The distance between two instances is calculated as follows: for the first instance (a1, a2, ... an); for the next instance (b1, b2, ... bn):

$$\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\cdots\cdots\cdots\cdots(a_n-b_n)^2}$$

The Euclidean distance formula has a major problem in that large values deluge the smaller ones. Heart disease records, for example, vary in cholesterol measurement from 100 to 190 and in age measure from 40 to 80. The cholesterol measurement is thus more influential than the age measurement. The solution to this problem is to normalize the continuous s properties such that they all have the same impact on the range among occurrences. Let us try to study KNN algorithm

As a rule, KNN is focused on continuous attributes, although it can also be applied to discrete attributes. The difference between the attribute values for two discrete instances a2, b2 is equal to one when the attribute values of those two instances are distinct, otherwise it is equal to zero.

D. Naïve Bayes algorithm

Naive Byes calculation depends on the idea of Bayes theorem. It calculates the independence of a feature. It also checks whether a feature is dependent on any other feature.

By the given table probability of hyperthyroidism present is O.6 and probability of hyperthyroidism absent is 0.4.
Through examination of symptoms if we find that the probability is >= 0.6 then hyperthyroidism is confirmed.

After feeding the datasets to the algorithm the disease is being predicted. Simultaneously the measures are also displayed for the predicted disease. This in turns makes sure that the patient is given proper advice.

IV IMPLEMENTATION

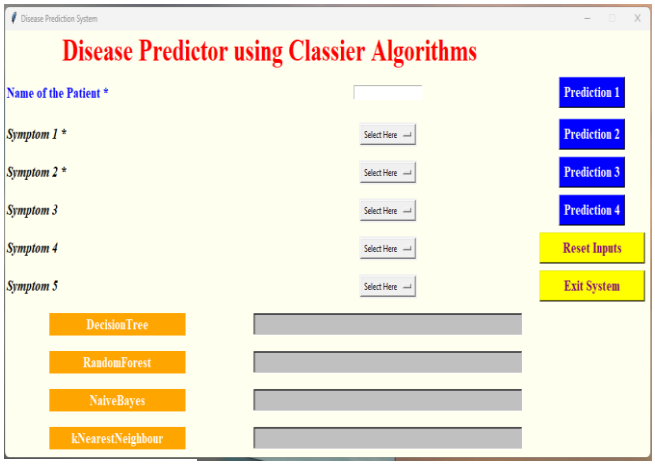


Fig 1: Disease Prediction GUI

To begin, the user must provide his/her name. In case the user is unable to input his/her name, a pop-up message will appear asking him/her to enter the name, as shown in Figure 2. User symptoms are taken into account in the GUI. You can choose up to five symptoms, and at least two are required. If not, you receive a pop-up, as shown in Figure 3.

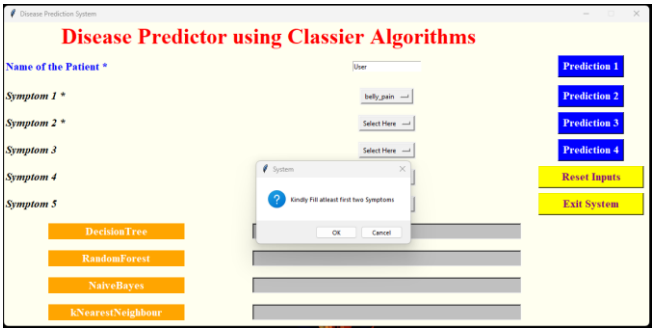


Fig 2: Pop-up to enter the name

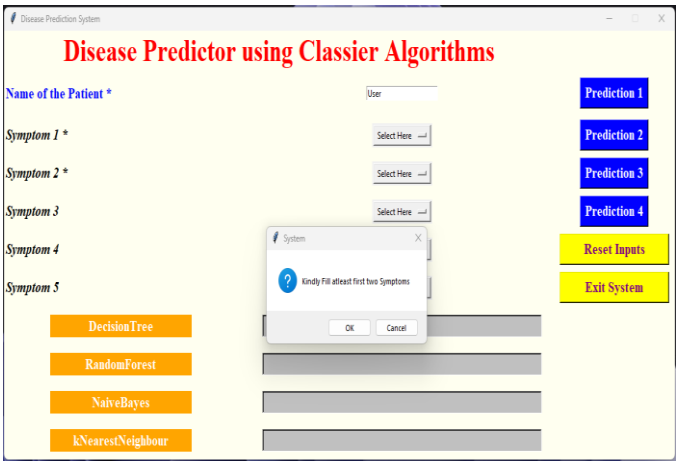


Fig 3: Pop-up to enter at least 2 symptoms

When the user enters the name and selects the symptoms they experience, through clicking on the *Prediction 1, Prediction 2, Prediction 3, Prediction 4* buttons, each algorithm will display the disease which it has predicted based on the symptoms. The disease predictions for the given symptoms is shown in Fig. 4. The new user will have the option to reset all the inputs if they want to know the disease for their symptoms, or will have the option to exit using the button *Exit System* as shown in Fig 5 if the existing user wishes to quit.

Fig 4 – Prediction

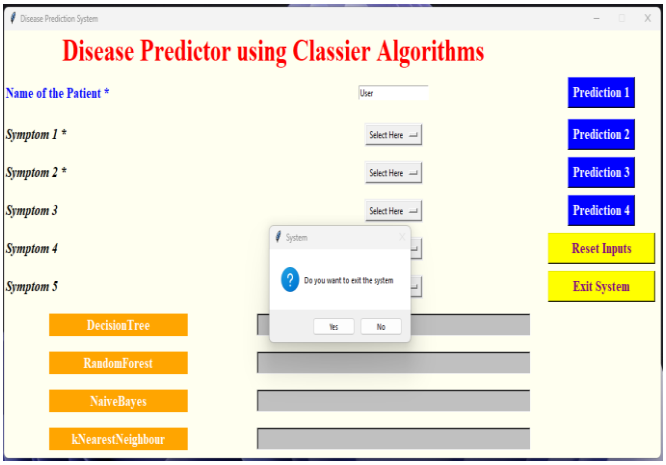


Fig 5: Exit system

V CONCLUSION

With regard to the application of machine learning to the medical sector, we can see that systems and methodologies have been developed through the development of simple and straightforward algorithms that are used to process sophisticated data. In this paper, we present a comprehensive comparison of three algorithms that were all able to yield a very good percent accuracy on a medical record. Artificial Intelligence will have a greater role to play in data analysis in the future as a result of the availability of huge amounts of data produced by modern technology and stored. Recent technologies have helped humans to lead better life. Using these techniques in health sector helped to make the decisions clear and faster.

VI REFERENCES

[1] Disease Prediction using Machine Learning Algorithms Sneha Grampurohit, Chetan Sagarnal 2021
[2] Web Based Disease Detection System Guided by Hasnuhana Mazumder Sayantan Saha, Argha Roy Chowdhuri, Anindita Dey and Sourav Haldar 2018
[3] Min Chen, Yixue Hao et.al "Disease Prediction by Machine Learning over big data from Healthcare Communities", IEEE[Access 2017]
[4] Palli Suryachandra, Prof.Venkata Subba Reddy,"Comparison of Machine Learning algorithms For Breast Cancer", IEEE 2013
[5] M. K. Obenshain, "Application of Data Mining Techniques to Healthcare Data," Infection Control and Hospital Epidemiology, 2004.
[6] V. A. Sitar-Taut et al., "Using machine learning algorithms in cardiovascular disease risk evaluation," Journal of Applied Computer Science and Mathematics, 2009.
[7] T. Porter and B. Green, "Identifying Diabetic Patients: A Data Mining Approach," Americas Conference on Information Systems, 2009.
[8] L. Li, T. H., Z. Wu, J. Gong , M. Gruidl, J. Zou, M. Tockman, and R. A. Clark, "Data mining techniques for cancer detection using serum proteomic profiling," Artificial Intelligence in Medicine, Elsevier, 2004
[9] R. Das, I. Turkoglu, and A. Sengur, Effective diagnosis of heart disease through neural networks ensembles, Expert Systems with Applications, Elsevier, pp. 7675–7680, 2009.
[10] UCI Machine Learning Repository: Pima Indians Diabetes",[online]Available: https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/
[11] N. Tazin, S. A. Sabab and M. T. Chowdhury, "Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique," 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec), Dhaka, 2016, pp. 1-6.
[12] B.Bharathi, S.Prince Mary (2019), Neural Computation based general disease prediction model, International journal of Recent Technology and Engineering, vol8(2),pp 5646-5649.
[13] Kamalesh M.D., Predictind the risk of diabetes mellitus to subpopulations using association rule mining, Proceedings of the International Conference of SoftComputing systems, Advances in Intelligent Systems and Computing col.397, Springer (2016)
[14] Revathy,B.Parvathavarthini,Shiny Caroline, Decision Theory, an Unprecedented Validation Scheme for Rough-Fuzzy Clustering, International Journal on Artificial Intelligence Tools, World Scientific Publishing Company ,Vol.25,No.2,2016
[15] Dwivedi AK. "Performance evaluation of different machine learning techniques for prediction of heart disease", Springer, Computer Applications and Mathematics. 2016
[16] Pahwa K, Kumar R. "Prediction of heart disease using hybrid technique for selecting features," 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura. 2017. p. 500–504. https://doi.org/10.1109/UPCON.2017.8251100.
[17] Indhumathi S, Vijaybaskar G. "WEB-BASED HEALTH CARE DETECTION USING NAIVE BAYES ALGORITHM." 2015.
[18] Saini R, Bindal N, Bansal P. "Classification of heart diseases from ECG signals using wavelet transform and kNN classifier," International Conference on Computing, Communication & Automation, Noida. 2015. p. 1208–1215. https://doi.org/10.1109/CCAA.2015.7148561.