iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** II **Month of publication:** February 2023

DOI: https://doi.org/10.22214/ijraset.2023.48954

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue II Feb 2023- Available at www.ijraset.com

Application of Machine Learning in Disease Prediction

Ayush Garg[1], Deepika Bansal[2]
[1, 2]*Department of IT, Maharaja Agrasen Institute of Technology, Delhi, India*

Abstract: ***Healthcare is a sector that is always changing. Healthcare professionals may find it challenging to stay current with the constant development of new technologies and treatments. As a result, the purpose of this research paper is to try and implement machine learning features in a specific system for health facilities. Knowing if we are ill at an early stage rather than finding out later is crucial. The entire process of treatment can be made much more effective if the disease is predicted ahead using specific machine learning algorithms as opposed to directly treating the patient. In this work, disease is predicted based on symptoms using machine learning. Machine learning algorithms like Naive Bayes, Decision Tree, Random Forest, and KNN are used to forecast the disease on the provided dataset. As you can see, there are numerous potential applications of machine learning in clinical care in the areas of patient data improvement, diagnosis, and treatment, cost reduction, and improved patient safety.***

I. INTRODUCTION

Over the past 20 years, the healthcare sector has undergone a great deal of change. As hospitals struggled to handle the massive influx of patients and public health officials worked to stop the spread of disease, the COVID-19 pandemic brought these difficulties to light. In order to accurately predict a person's disease based on the symptoms they exhibit, this article aims to implement a reliable machine learning model. An essential component of treatment is the ability to diagnose disease based on its symptoms. In my project, I have attempted to correctly predict a disease by considering the patient's symptoms. I achieved an accuracy of 94% to 95% by using 4 different algorithms for this. Future medical treatment could greatly benefit from a system like this. To make using the system easier, I also created an interactive interface.

II. LITERATURE REVIEW

A lot of research work has been done to predict diseases based on the symptoms an individual exhibits using machine learning algorithms. Monto et al. [1] proposed a statistical model to predict whether a patient had influenza or not. They included 3,744 unvaccinated adults and adolescents with influenza who had fever and at least 2 other flu symptoms. Of the 3,744, 2,470 were confirmed to have the flu.

Based on this data, their model provided an accuracy of 79%. Sreevalli et al. [2] used a random forest machine learning algorithm to predict disease based on symptoms. The system brought low time consumption and minimal costs for disease prediction. The algorithm resulted in an accuracy of 84.2%. Langbehn et al. [3] to detect Alzheimer's disease. Data for 29 adults were used for ML algorithm training purposes. They developed classification models to detect reliable absolute changes in scores using the SmoteBOOST and wRACOG algorithms. Various ML techniques, such as artificial neural networks (ANNs), Bayesian networks (BNs), support vector machines (SVMs), and decision trees (DTs), have been widely used in cancer research to develop predictive models, resulting in efficient and accurate decision making [4].

Karayilan et al. [5] proposed a heart disease prediction system that uses an artificial neural network backpropagation algorithm. The 13 clinical symptoms were used as input for the neural network and then the neural network was trained using a backpropagation algorithm to predict the absence or presence of heart disease with 95% accuracy. Various machine learning algorithms have been streamlined for effective chronic disease outbreak prediction by Chen et al. [6].

III. DATABASE COLLECTION

A study from Columbia University was conducted at New York Presbyterian Hospital in 2004. The data for this project was gathered from that study. Below is a link to the dataset: - http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html

A. Gathering the Data

Any machine learning problem must start with data preparation. We'll be using data from a 2004 study by Columbia University that was conducted at New York Presbyterian Hospital. A training file and a testing file are both CSV files that make up this dataset. 42 different types of diseases can be predicted using the 132 parameters in this dataset.

B. Cleaning is the most important stage in *Cleaning the Data*

a machine learning project. Our machine learning model's quality is based on the effectiveness of our data. Consequently, the data must always be cleaned before being fed to the model for training. A label encoder is used to convert the string-type prognosis column to numerical form. The dataset we are using has just number columns.

C. Dividing Input Data into Training and Testing Sets

To assess how well our machine learning model works, we must divide a dataset into train and test sets. We'll divide the data into an 80:20 structure, meaning that 80% of the dataset will be used to train the model and 20% of the data will be used to assess how well the model performed.

IV. PROPOSED SYSTEM

A. Creating the Classification Model

The data can be used to train a machine learning model once it has been collected and cleaned. The Decision Tree Classifier, Naive Bayes Classifier, K Nearest Neighbor, and Random Forest Classifier will all be trained using the cleaned data. To assess the model's quality, a confusion matrix will be used.

B. Compiling the Model

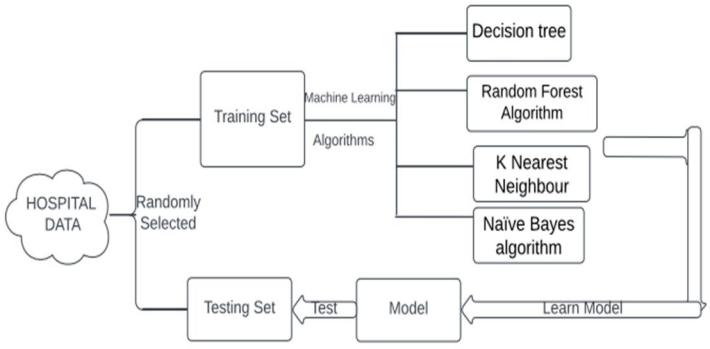
The model must first be defined before we can compile it. The performance of various algorithms was then evaluated based on the accuracy scores assigned to each of the different algorithms used to carry out the various procedures. Only a small portion of the dataset's variables are relevant for creating the machine learning model; the remaining features are either redundant or unimportant. If we add all these redundant and meaningless features to the dataset, the model's overall accuracy and performance may decline.

C. Fitting the Model

It is determined by how well a machine learning model fits the data when it generalizes data that is comparable to the data it was trained with. A good model fit is an accurate approximation of the outcome when given uncertain inputs. As a result, we can forecast data that the data model has never seen in the future. Let's imagine we want to assess how well our machine learning model absorbs new data and changes in response to it.

D. Evaluating the Model

Following the completion of the step, the model can be assessed using various criteria, including accuracy, recall, precision, and f1 score. After the four models have been trained, we will combine their predictions to predict the disease from the input symptoms. As a result, our prediction is stronger and more precise overall. Finally, we will define a function that takes a set of symptoms as input, infers the disease from those symptoms using trained models, and outputs the prediction using the Tkinter GUI.



V. MODELS

In our project, four different types of disease prediction models are present. These models are:-

1) Decision tree
2) Random forest
3) Gaussian Naïve Bayes
4) K Nearest Neighbor (KNN)

A. *Decision Tree*

1) The decision tree algorithm is a type of supervised learning. Both classification and regression issues can be resolved using them.
2) Decision trees use a tree representation to solve problems, with each leaf node denoting a class label and the inside nodes of the tree housing attribute representations.
3) Using the decision tree, we may represent any Boolean function on discrete characteristics.

Our project's initial prediction strategy was a decision tree. It provides us with a 93 percent accuracy rate.

B. *Random Forest Algorithm*

1) Using decision trees, classification, regression, and other tasks are performed using the supervised machine learning method known as the Random Forest or Random Decision Forest.
2) The Random Forest classifier creates a set of decision trees from a subset of training data that is randomly selected. It merely comprises of a collection of decision trees (DT) drawn from a randomly selected subset of the training set, with the final prediction being determined using those decision trees.

In this project, we used a random forest classifier with samples of test data, and the result is 94 percent accurate.

C. *K Nearest Neighbour*

1) K-Nearest Neighbor, one of the simplest machine learning algorithms, employs the supervised learning approach.
2) The K-NN algorithm places the new instance in the category that is most like the existing categories on the assumption that the new case and the old cases are comparable.
3) After all the previous data has been recorded, a new data point is categorized using the K-NN algorithm based on similarity. This means that using the K-NN approach, new data may be sorted into the appropriate category rapidly and accurately.
4) Although the K-NN technique can be used to solve classification and regression problems, classification challenges are the ones for which it is most usually applied.

In this research, we employed a KNN classifier using samples of test data, and the outcome is approximately 95% accurate.

D. *Naïve Bayes algorithm*

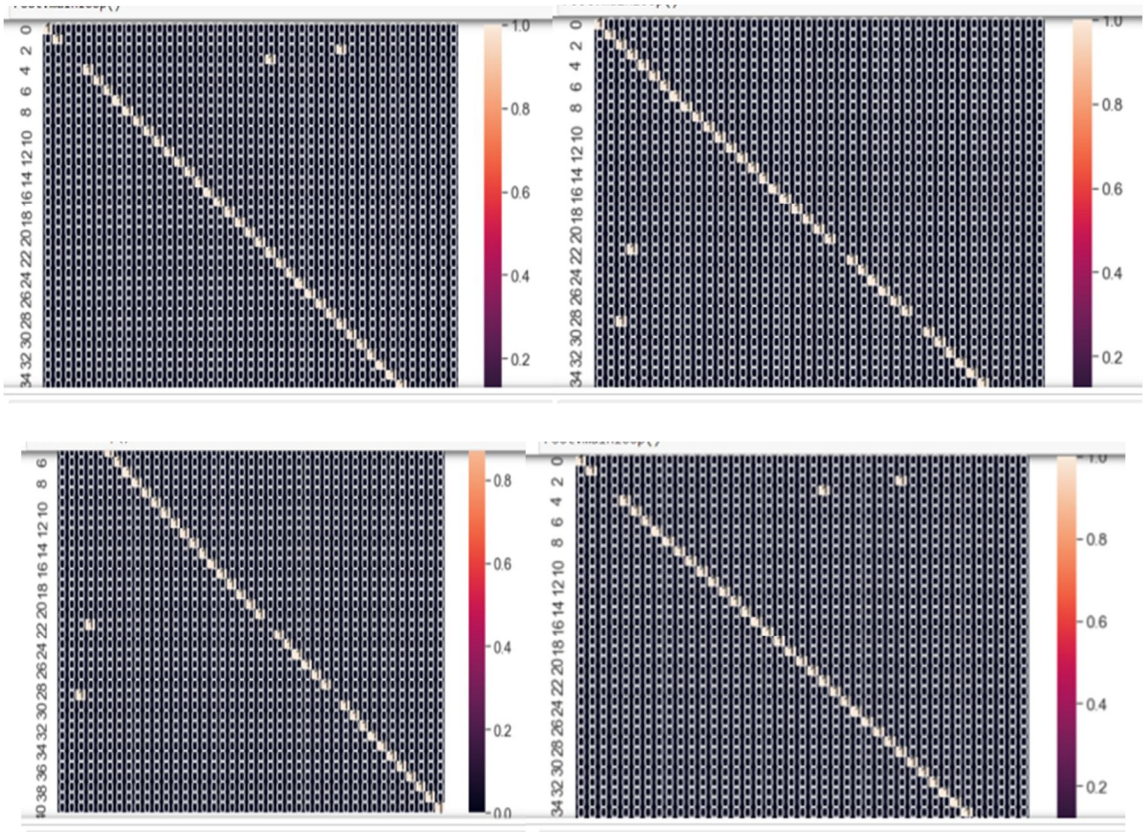
1) Based on the Bayes theorem, the Naive Bayes algorithm is a supervised learning technique for classification problems.
2) One of the simplest and most effective classification methods for building rapid machine learning models that can make precise predictions is the Naive Bayes Classifier.
3) Because it is a probabilistic classifier, predictions are dependent on the likelihood that an object actually exists.

In our project, we employed the naive Bayes algorithm to obtain a forecast that was 95% correct.

VI. RESULTS

Model	Accuracy	Recall_Score	Precision_Score	F1_Score
Decision Tree	0.93454	0.93909	0.91496	0.92213
Random Forest	0.93454	0.93909	0.91496	0.92213
Naive Bayes	0.94554	0.93815	0.91545	0.92235
K Nearest Neighbour	0.95655	0.93909	0.91618	0.92320

A. Confusion Matrix



VII. CONCLUSION

Our goal was to build a system that, given a set of symptoms, could diagnose diseases and predict their occurrence. A system like this could lessen the rush at hospital OPDs and ease the burden on the medical staff. Using 4 different algorithms, we were successful in developing such a system. We managed a 95 percent accuracy rate on average. A system like this is generally capable of performing the task.

While developing this system, it is also possible to add a method of storing user inputted data in a database that can be used later to aid in the development of future iterations of the system. Our system also has a user-friendly interface. Additionally, the data gathered, and outcomes obtained are represented visually in a variety of ways.

VIII. FUTURE SCOPE

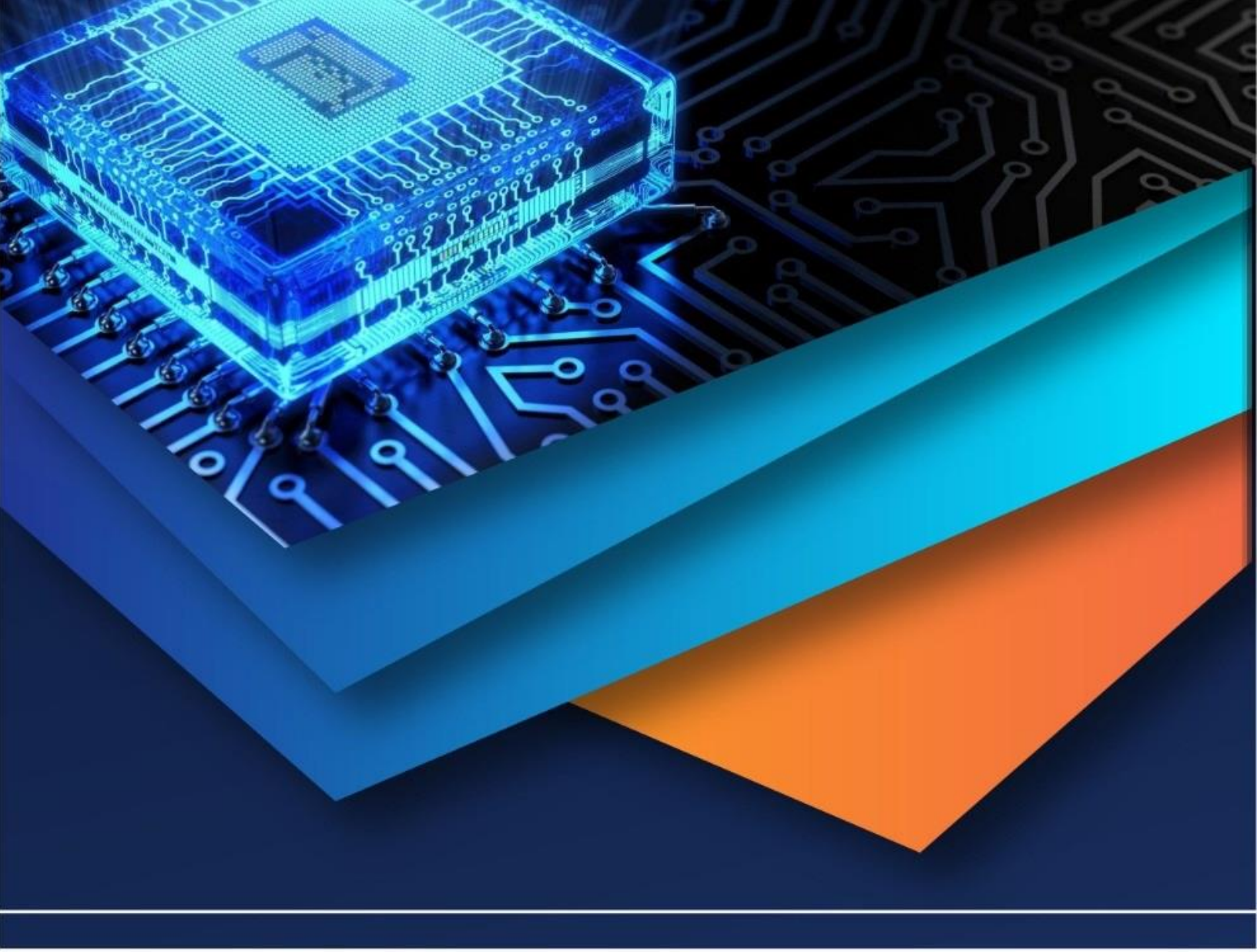
Data-driven decisions have dominated the 21st century. According to conventional wisdom, sectors or industries that produce more data will expand more quickly, and businesses that use that data to make key decisions may be able to stay ahead of the curve. The healthcare sector is among the top ones when it comes to producing a lot of data. Thanks to a variety of fresh approaches to data collection, like sensor-generated data. What if this data could be used to improve healthcare services while lowering costs and improving patient satisfaction? It's possible by utilizing machine learning techniques.

Healthcare professionals who effectively implement machine learning are able to make better decisions, spot trends and innovations, and increase the effectiveness of research and clinical trials. The difficulty is in gathering this data and effectively using it for analysis, prediction, and treatment as healthcare generates a lot of it. Let's examine how machine learning can handle this problem. Instead of treating a patient after a diagnosis, the modern healthcare approach focuses on early disease intervention and disease prevention. A risk calculator is typically used by medical professionals to determine the likelihood of developing a disease. These calculators estimate the likelihood of contracting a particular disease using basic data from demographics, health status, daily activities, and more. Equation-based mathematical techniques and tools are used to perform these calculations. When using a similar equation-based methodology, the problem is the low accuracy rate.

Data science and machine learning are driving medicine into a new area. Events and procedures that were solely an idea a few years ago have now shifted into reality. Now we are living in the era of machine learning, where algorithms can encourage us in stopping and handling diseases by examining our data and helping doctors to make better decisions.

REFERENCES

[1] A.S. Monto, S. Gravenstein, M. Elliott, M. Colopy, J. Schweinle, Clinical signs and symptoms predicting influenza infection, Archives of internal medicine 160(21), 3243 (2019)
[2] R.D.H.D.P. Sreevalli, K.P.M. Asia, Prediction of diseases using random forest classification algorithm
[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 (2018)
[4] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and structural biotechnology journal 13, 8 (2015)
[5] T. Karayılan, O. Kılı¸c, in 2017 International Conference on Computer Science and Engineering (UBMK) (IEEE, 2017), pp. 719–723
[6] M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, Disease prediction by machine learning over big data from healthcare communities, Ieee Access **5**, 8869 (2017)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY