IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 3, March 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303247 |

Disease Prediction Using Symptoms in Machine Learning

[1]Mrs.T. Neetha, [2]U. Lokesh Reddy, [3]G. Bhargavi, [4]K. Sanjana Reddy

[1]Assistant Professor, Department of Artificial Intelligence, Anurag University, Hyderabad, India

[2,3,4] U.G. Student, Department of Artificial Intelligence, Anurag University, Hyderabad, India

ABSTRACT*:* Disease prediction using symptoms is a critical area of research in healthcare aimed at improving early diagnosis and treatment planning. Leveraging machine learning techniques, this study presents an approach to predict diseases based on symptom data. A dataset containing symptoms and corresponding diagnosed diseases is utilized to train and validate various machine learning models, including decision trees, support vector machines, and neural networks. Feature selection methods are employed to identify the most relevant symptoms for accurate prediction. The models are evaluated using metrics such as accuracy, precision, recall, and F1-score to assess their performance in predicting diseases. The results demonstrate promising predictive capabilities of machine learning models, showcasing the potential for early disease detection and aiding healthcare professionals in making informed decisions for patient care. Further refinement and integration of additional data sources and advanced machine learning algorithms can enhance the accuracy and applicability of disease prediction systems, ultimately contributing to proactive and efficient healthcare management.

KEYWORDS*:* Machine Learning, Decision Trees, Support Vector Machines, Accuracy, Neural Networks, Precision, Recall, F1-score.

I. INTRODUCTION

When anyone is currently afflicted with an illness, they must see a doctor, which is both time consuming and costly. It can also be difficult for the user if they are not near to doctors and hospitals because the illness cannot be identified. So, if the above procedure can be done using an automated software that saves time and money, it could be better for the patient, making the process go more smoothly.

Disease prediction using symptoms through machine learning approaches has emerged as a pivotal research area within the realm of healthcare informatics. Accurate and timely prediction of diseases based on a patient's symptoms plays a crucial role in early intervention, treatment planning, and improving overall healthcare outcomes. Machine learning offers a powerful framework for analyzing large volumes of data and extracting meaningful patterns, making it an ideal tool for disease prediction using symptom-based.

This study explores the application of various machine learning algorithms to predict diseases based on symptoms, aiming to enhance diagnostic efficiency and provide healthcare practitioners with valuable insights for delivering timely and personalized medical interventions. By utilizing an extensive dataset that includes information on symptoms and associated medical conditions, we apply machine learning algorithms and feature selection methods to develop a prediction system that works well. By assessing these models using pertinent performance indicators, new opportunities for improving patient outcomes and healthcare management are created, and the promise of machine learning in illness prediction is shown.

II. LITERATURE SURVEY

Numerous research works have been carried out for the prediction of the diseases based on the symptoms shown by an individual using machine learning algorithms.

Monto et al. designed a statistical model to predict whether a patient had influenza or not. They included 3744 unvaccinated adults and adolescent patients of influenza who had fever and at least 2 other symptoms of influenza. Out of 3744, 2470 were confirmed to have influenza by the laboratory. Based on this data, their model gave an accuracy of 79 %.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303247 |

Sreevalli et al. used the random forest machine-learning algorithm to predict the disease based on the symptoms. The system resulted in low time consumption and minimal cost for the prediction of diseases. The algorithm resulted in an accuracy of 84.2 %. Various tools were developed by Langbehn et al. to detect Alzheimer's disease. Data for 29 adults were used for the training purpose of the ML algorithm. They had developed classification models to detect reliable absolute changes in the scores with the help of SmoteBOOST and wRACOG algorithms. A variety of ML techniques such as artificial neural networks (ANNs), bayesian networks (BNs), support vector machines (SVMs) and decision trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making

Karayilan et al. proposed a heart disease prediction system that uses the artificial neural network backpropagation algorithm. 13 clinical features were used as input for the neural network and then the neural network was trained with the backpropagation algorithm to predict absence or presence of heart disease with an accuracy of 95 %. Various machine learning algorithms were streamlined for the effective prediction of a chronic disease outbreak by Chen et al. . The data collected for the training purpose was incomplete. To overcome this, a latent factor model was used. A new convolutional neural network-based multimodal disease risk prediction (CNN-MDRP) was structured. The algorithm reached an accuracy of around 94.8 %.

Chae et al. used 4 different deep learning models namely deep neural networks (DNN), long short term memory (LSTM), ordinary least squares (OLS), an autoregressive integrated moving average (ARIMA) for monitoring 80 infectious diseases in 6 groups. Of all the models used, DNN and LSTM models had a better performance. The DNN model performed better in terms of average performance and the LSTM model gave close predictions when occurrences were large. Haq et al. used a database that contained information about patients having any heart disease. They extracted features using three selection algorithms which are relief, minimum redundancy, and maximum relevance (mRMR), and least absolute shrinkage and selection operator which was cross-verified by the K-fold method. The extracted features were sent to 6 different machine learning algorithms and then it was classified based on the presence or absence of heart disease. An effective heart disease prediction system was developed by Mohan et al. .They achieved an accuracy level of 88.4 % through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM). Maniruzzaman et al. classified the diabetes disease using ML algorithms. Logistic regression (LR) was used to identify the risk factors for diabetes disease. The overall accuracy of the ML-based system was 90.62 %..

III. PROBLEM STATEMENT

The challenge lies in accurately predicting diseases based on symptoms, a task crucial for early diagnosis and effective treatment. Conventional diagnostic methods are often time-consuming and costly, necessitating the exploration of machine learning techniques to develop a predictive model that aids in timely disease detection and enhances healthcare decision-making.

3.1 Existing Systems

Existing disease prediction models analyze the symptoms someone has and use computer patterns to guess which illness they might have. These models help doctors identify problems early and suggest better treatments for patients.

3.2 Proposed System

Disease prediction models now extend beyond risk assessment, incorporating personalized dietary recommendations based on the patients disease. By combining patient data with medical knowledge, these models offer tailored medical guidance to mitigate disease risks. This holistic approach enhances preventive healthcare, empowering individuals to make informed choices that align health.

IV. DATASET

Dataset consists of total 133 columns out of which 132 columns represent the symptoms with 42 different types of diseases which can be predicted and the last column is the prognosis. In our dataset all the columns are numerical. The dataset is divided into Two classes named as Training Dataset and Testing dataset which are two separate folders.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303247 |

V. PROPOSED METHODOLOGY

Step 1 - First we will collect the datasets of symptoms and their functional problem in the body.
Step 2 - Then we will collect the information that will associate the symptoms to possible diseases thus related disease information will be collected.
Step 3 - Then we will get the symptoms as input from the patient and process it by SVM, Naive Bayes, Random Forest.
Step 4 - After that these algorithms predicts the diseases that may be possible for those acquired symptoms.
Step 5 - Then the system will show the diagnosis and also recommend medicines.

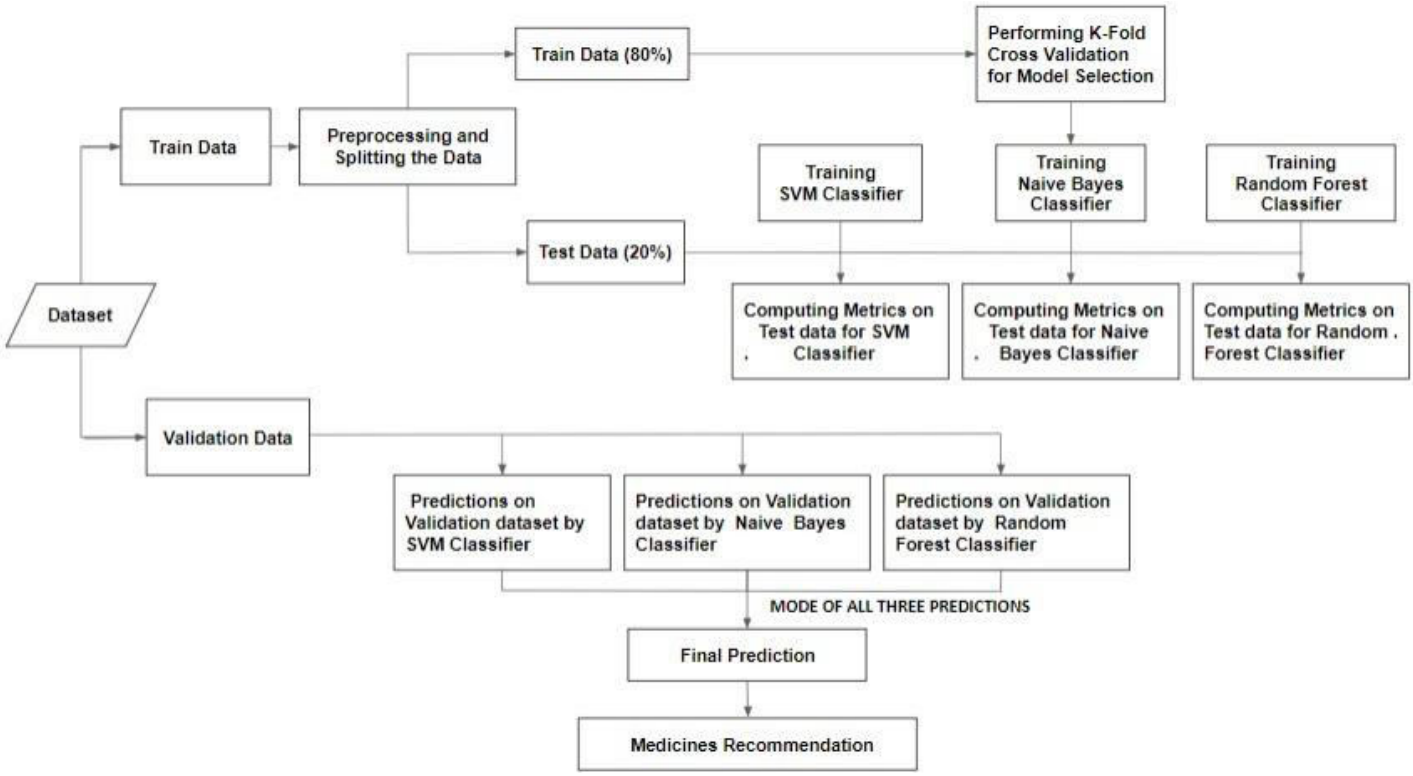

Figure : Flow Chart

5.1. Data Collection

Data collection or gathering is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. Data collection is a research component in all study fields, including physical and social sciences, humanities, and business. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same. The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed. Data collection and validation consists of four steps when it involves taking a census and seven steps when it involves sampling. The data is taken from the kaggle it contains 132 parameters on which 42 different types of diseases can be predicted.

5.2. Data Preprocessing and cleaning

Information preprocessing can allude to control or dropping of information some time recently it is utilized in arrange to guarantee or upgrade performance and is an imperative step within the information mining prepare. The express "waste in, waste out" is especially appropriate to information mining and machine learning ventures. Data-gathering strategies are frequently freely controlled, coming about in out-of-range values (e.g., Pay:−100),

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303247 |

incomprehensible information combinations (e.g., Sex: Male, Pregnant :Yes), and lost values, etc. The Collected information may contain a few the invalid esteem. When the dataset was being arranged some of the values within the parameters may be misplaced so when we are attempting to fit our show on thedataset which have the invalid values at that point our demonstrate will toss an mistake saying that the a few of the values which are show in our collected dataset are null(empty) so we must clear those specific exchange which means we must drop those specific exchanges in arrange to dodge the mistakes.

5.3. Splitting The Dataset

Every machine learning problem starts with information. ML models without proper information are like bodies without souls. But gathering data isn't as much of a problem in the "big data" world of today. Whether on purpose or accidentally, we produce enormous databases every day. That being said, having access to overflow information does not resolve the problem. In addition to requiring vast volumes of data to be supported, we also need to ensure the quality of the data in order for ML models to produce meaningful results. Even while deciphering raw data into meaningful insights is a skill in and of itself, requiring exceptional spatial intelligence and design skills (in rare instances), high-quality data remains useless unless it is used correctly. The main problem that ML/DL experts face is dividing up the data into training and testing sets. Although at first glance it seems like a simple problem, its complexity may be measured, so to speak, by delving deeply into it. Poor planning and testing sets may have unexpected effects on the demonstrate's yield. It may cause the data to be either overfitted or underfitted, and our show might end up providing biassed results.

5.4 Applying Machine Learning Algorithms

After splitting the data, we will be using K-Fold cross-validation to evaluate the machine-learning models. We will be using Support Vector Classifier, Gaussian Naive Bayes Classifier, and Random Forest Classifier for cross-validation.

5.5 Calculate Performance Parameters

This step consists of calculating different performance measures such as recall, precision, accuracy etc the model with themost acceptable is the one which is having good accuracy and good recall and good precision is considered as a good model i.e The next step after implementing a machine learning algorithm is to find out how effective is the model based on metric and datasets. Different performance metrics are used to evaluate different Machine Learning Algorithms. For example a classifier used to distinguish between images of different objects; we can use classification performance metrics such as, Log-Loss, Average Accuracy, AUC, etc. If the machine learning model is trying to predict a stock price, then RMSE (root mean squared error) can beused to calculate the efficiency of the model. Another example of metric for evaluation of machine learning algorithms is precisionrecall or NDCG, which can be used for sorting algorithms primarily used by search engines. Therefore, we see that different metrics are required to measure the efficiency of different algorithms, also depending upon the dataset at hand.

```
============================================================
SVC
Scores: [1. 1. 0. 1. 1. 1. 1. 1. 1. 1.]
Mean Score: 0.9
============================================================
Gaussian NB
Scores: [1. 1. 1. 1. 0. 1. 1. 1. 0. 1.]
Mean Score: 0.8
============================================================
Random Forest
Scores: [0. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Mean Score: 0.9
```

Figure: Metric Scores

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303247 |

VI. RESULTS

After an extensive training process on a large dataset, the model has achieved impressive results in terms of accuracy. The model's superior performance in both training and testing phases validates its effectiveness as a powerful classifier, capable of accurately categorizing data into appropriate. The model's consistent and impressive results highlight its reliability and suitability for real-world scenarios, making it a promising choice for diverse machine learning and artificial intelligence applications.

6.1 SVM

Table depicting the accuracy of SVM applied on datasets of sizes 2600 and 3500.

DATASET	TRAINING	TESTING	ACCURACY
2600	1950	650	85
3500	2625	875	88

The above table displays the results after applying the **SVM** model for the datasets consisting of 2600 and 3500 rows. On further dividing the 3500 rows dataset into training and testing with a ratio of 75:25, we achieved an accuracy of 88.00%. Similarly on dividing the 2600 rows dataset in the ratio of 75:25, we achieved an accuracy of 85.00%. The maximum accuracy was encountered in the case of 3500 rows dataset which is 88.00%.

6.2 NAÏVE BAYES

Table depicting the accuracy of NAÏVE BAYES applied on a dataset of 2600 and 3500.

DATASET	TRAINING	TESTING	ACCURACY
2600	1950	650	80
3500	2625	875	84

The table displays the results after applying the NAÏVE BAYES model for a dataset consisting of 2600 rows and 3500 rows. The datasets were designed in a similar way as that of the previous model. On further dividing the 3500 rows dataset into training and testing with a ratio of 75:25, we achieved an accuracy of 84.00%. Similarly on dividing the 2600 rows dataset in the ratio of 75:25, we achieved an accuracy of 80.00%. The maximum accuracy was encountered in the case of 3500 rows dataset which is 84.00%.

6.3 RANDOM FOREST

Table depicting the accuracy of RANDOM FOREST applied on a dataset of 2600 and 3500.

DATASET	TRAINING	TESTING	ACCURACY
2600	1950	650	90
3500	2625	875	92

Table 3 displays the results after applying the RANDOM FOREST model for a dataset consisting of 2600 rows and 3500 rows. The datasets were designed in a similar way as that of the previous model. On further dividing the 3500 rows dataset into training and testing with a ratio of 75:25, we achieved an accuracy of 92.00%. Similarly on dividing

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303247 |

the 2600 rows dataset in the ratio of 75:25, we achieved an accuracy of 90.00%. The maximum accuracy was encountered in the case of 3500 rows dataset which is 92.00%.

On comparing the results of all the three models, we can observe that the Random Forest model has outperformed SVM and Naïve Bayes models in the case of all the datasets. The maximum accuracy achieved through SVM and Naïve Bayes is 88.00% and 84.00%, where the accuracy achieved from the Random Forest model is 92.00%.

VII. CONCLUSION

In this project, the model was able to predict diseases by classifying them with the help of symptoms, which are given as input to the model. In order to bring out the maximum efficiency of the model, three different approaches are used. They are SVM, Naïve Bayes, and Random Forest. Out of these three algorithms, Random Forest predicted most accurately.

ACKNOWLEDGMENT

We want to express our deep-felt gratitude and sincere thanks to our guide Mrs.T.Neetha, Assistant Professor, Department of AI, Anurag University, for her skillful guidance, timely suggestions, and encouragement in completing this project.We want to express our profound gratitude to all for having helped us in achieving this dissertation. Finally, we would like to express our heartfelt thanks to our parents, who were very financially and mentally supportive and for their encouragement to achieve our goals.

REFERENCES

1. Shuai Zheng, Zhenfeng Zhu, Zhizhe Liu, Zhenyu Guo, Yang Liu, Yuchen Yang, Yao Zhao, “Multi-Modal Graph Learning for Disease Prediction”, IEEE Transactions on Medical Imaging, vol. 41, July 2021.
2. Jianliang Gao, Ling Tian, Jianxin Wang, “Similar Disease Prediction with Heterogeneous Disease Information Networks”, IEEE Transactions on NanoBioscience, 2020.
3. Md. Ehtisham Farooqui, Dr. Jameel Ahmad, “Disease Prediction System using Support Vector Machine and Multilinear Regression”, Volume- 8, July- 2020
4. Mohd. Nadeem Khan and Ankita Srivastava, “Disease Prediction using Machine Learning”, June 2023.
5. https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning
6. Sai Sriram Krishna Parimi ,Y.Snehith Reddy, “Prediction Of Multiple Diseases Using Machine Learning Techniques”, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details