Dr Rajkumar, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

# Disease Prediction from Symptoms using Machine Learning

**Assistant Professor Dr Rajkumar, Mehreen, Abiha kazmi, Km Ilma, Apeksha Kaushik, Kritika**
Quantum University Roorkee

**Abstract- The goal of this project is to develop a smart system that can forecast illnesses like typhoid, dengue, and malaria by analyzing patient-reported symptoms. It uses three different machine learning algorithms—Random Forest, Decision Tree, and Support Vector Machine—to identify diseases from a predefined list of symptoms and the corresponding diagnosis. Every algorithm has undergone rigorous training and testing to guarantee its precision and dependability in forecasting. When it comes to disease prediction, the Random Forest method outperforms the Decision Tree and Support Vector Machine models. By accurately identifying the relevant illness based on the user's symptoms, it consistently yields reliable results, making it the most effective model in our system.**

**Keywords— Machine Learning, Disease Prediction, SVM, Random Forest, Decision Tree, Classification**

## I. INTRODUCTION

### 1. Motivation

Using enormous volumes of historical data to enhance healthcare outcomes, machine learning (ML) has become a key technique in disease prediction. With the use of both organized and unstructured medical data, computers can now recognize trends and forecast outcomes. An ML model is created in two main stages: testing, which assesses the model's accuracy using fresh data, and training, which improves the model's predicting abilities using a particular dataset. ML algorithms analyze patient symptoms in the field of disease prediction, helping medical practitioners make more accurate and effective diagnosis.[1]

The speed and accuracy with which machine learning can process massive amounts of data highlights its significance in the healthcare industry. Healthcare data has grown more complex and abundant as data gathering techniques have improved, frequently incorporating both unstructured text and structured information. ML models can provide timely insights that are necessary for efficient decision-making by analyzing this data. This feature ultimately improves healthcare services by empowering doctors to make well-informed diagnoses and customize treatments to each patient's particular needs.[2]

Several machine learning techniques, including Classifiers SVM, RF, and DT, which are renowned for their capacity to analyze medical data and produce accurate disease predictions, are used in our work. The use of machine learning (ML) in healthcare simplifies patient data analysis, which promotes proactive monitoring and individualized treatment plans in addition to helping with diagnosis. The promise for better patient care and improved healthcare delivery becomes more attainable as the healthcare sector employs these advanced ML approaches more frequently.

### 2. Aim

The purpose of this study is to assess how supervised machine learning (ML) methods can improve healthcare by enabling prompt and precise disease diagnosis. This study focuses on investigating studies that use several supervised machine learning models for every disease recognition task. This Because evaluating the performance of a single algorithm across many study situations can create bias and produce inaccurate results, methodology ensures more comprehensiveness and accuracy. We'll look at a number of diseases, with a focus on dengue, typhoid, and malaria. Our approach evaluates the effectiveness of several techniques, such as DT, RF, and SVM, in predicting diseases including dengue, malaria, and typhoid, based on reported symptoms. The goal of this literature study is to identify the best machine learning models for each specific illness.

By using machine learning models to forecast diseases like dengue, malaria, and typhoid, this initiative has substantial benefits for the healthcare industry. By providing effective, precise, and easily available diagnostic support, it helps to modernize and enhance healthcare services. By providing customized treatment plans based on predictive data, all models have the potential to improve the disorder process and establish the foundation for developments in personalized healthcare. All things considered, incorporating machine learning into disease prediction is a critical step toward improved health outcomes and more effective healthcare delivery systems.

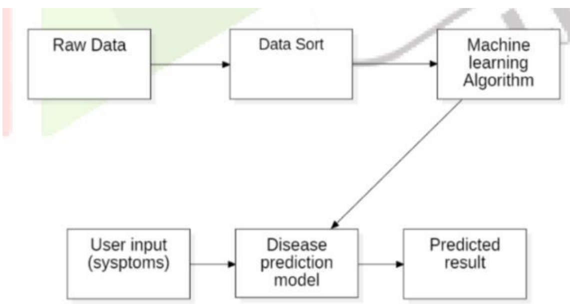


Figure 1. Block Diagram Proposed Architecture

## II. LITERATURE REVIEW

This project explores the application of machine learning to develop a system for predicting diseases including dengue, malaria, and typhoid in addition to include a module for hospital referrals. A number of classification methods, such as the Random Forest Classifier, Support Vector Machine, and Decision Tree, are used to predict the probability of an individual's illness based on their reported symptoms. The best kind of hospital for a consultation is then recommended. With an interactive front-end interface that links to a server, the system is intended for end users. This invention could have a big impact on how medicine is practiced in the future. But the intrinsic complexity and variety of illnesses may result in issues with accuracy and possible biases in the data used to train the algorithm.[3]

The study "Disease Prediction using Machine Learning" investigates how patient symptoms might be used to predict illnesses using machine learning techniques. It explains how models like Random Forest, Support Vector Machine (SVM), and Decision Trees can be used to diagnose illnesses like dengue, malaria, and typhoid. It also highlights how important precise illness prediction is for the medical industry.[4]

The Random Forest model uses an ensemble learning technique, mixing several decision trees to improve prediction accuracy in illness detection, whereas the Decision Tree model uses the CART algorithm to conduct classification. Additionally, SVM is used to maximize the separation between several classes in order to classify diseases. This project's main goals are to reliably forecast diseases based on user-provided symptoms, assess how well these categorization methods work, and examine how changes in symptom patterns impact the results of disease predictions.[5]

The researchers came to the conclusion that the suggested model showed improved accuracy, which they attributed to its ability to identify intricate nonlinear interactions in the feature space. Additionally, the Decision Tree (DT) successfully highlights important characteristics, providing a deeper understanding of the illness and enabling accurate forecasts for complex disorders.[6]

This paper presents a paradigm for early disease prediction using an ensemble model that incorporates Support Vector Machines (SVM). It shows how effective this approach is for a variety of illnesses, some of which may be of special interest to you. The study provides insightful viewpoints on the application of machine learning in the medical field for the early detection of various diseases, emphasizing the importance of accurate predictions and timely interventions to improve patient outcomes.[7]

In the traditional method of diagnosis, a patient must usually see a doctor, go through a number of tests, and then receive a diagnosis. It could take a long time to complete this process. In order to speed up the initial stage of disease identification that relies on user input, the suggested project presents an automated method for disease prediction. Users can choose to receive an estimate or prognosis of their possible disease depending on the information they offer when interacting with the chatbot thanks to the way the system is set up. However, it is important to recognize that when a small number of symptoms are recorded, prediction accuracy may decrease. [8] The accuracy rate is evaluated in this study using the Support Vector Machine approach. Metrics including accuracy, specificity, and sensitivity are used to assess the system's performance. The percentage of the sample data that is successfully and erroneously classified is displayed. This study looks at a number of models, including SVM, Random Forest, and Decision Tree. This system uses a reliable multi-process strategy that combines clustering techniques with the decision tree algorithm.

**Dataset Description**

The dataset utilized in this study includes medical data based on symptoms for machine learning-based disease prediction. The dataset, which was created in CSV format, contains a variety of illnesses and the symptoms that go along with them. Every entry in the dataset corresponds to an illness and a collection of symptoms that patients have reported.

Diseases like dengue, malaria, typhoid, stomach infections, skin infections, and pimples are included in the dataset. Fever, headache, nausea, vomiting, exhaustion, redness, swelling, itching, dehydration, and stomach discomfort are just a few of the symptoms. During preprocessing, these symptoms are first saved in text format and then transformed into a machine-readable format.

Before the machine learning models were trained, the dataset was cleaned and arranged. Preprocessing addressed conflicting symptom entries and missing values. To transform categorical symptom data into numerical values appropriate for machine learning algorithms, feature encoding techniques were used.

Three supervised machine learning methods were trained and tested using the provided dataset:

- Random Forest
- Decision Tree
- Support Vector Machine (SVM)

Because the quality and variety of symptoms directly impact model performance, the dataset is crucial to increasing prediction accuracy. To assess the effectiveness and dependability of the prediction models, the dataset was split into training and testing sets.

**Dataset Features**

- Disease Name
- Symptoms
- Prognosis/Prediction Label

**Advantages of Dataset**

- Contains multiple disease categories
- Easy to preprocess and train
- Suitable for classification algorithms
- Helps in early disease prediction

## III. PROPOSED METHODOLOGY

Our project employs a systematic methodology as follows:

**Motivation:** A comprehensive dataset that includes disease classifications and symptoms is used, as shown in the "prognosis" column. This dataset includes a variety of symptoms linked to illnesses including typhoid, dengue, and malaria, such as fever, headache, nausea, and muscle soreness.

**Data Preparation:** The cleaning and preparation of the dataset is the main focus of the next stage of our research. This procedure involves filling in any gaps in the symptom data and preserving the integrity and consistency of the dataset. To enhance the dataset and ensure effective training of the machine learning models for disease prediction, we will implement preprocessing approaches including encoding categorical symptoms and standardizing the input characteristics.

**Feature Selection:** Examine the dataset to identify the symptoms linked to the diseases in question, particularly typhoid, dengue, and malaria. To highlight the symptoms that are most typical of the three diseases under consideration, use feature engineering techniques.

**Model Selection**

- Random Forest Classifier: Uses a group of decision trees to improve classification accuracy and reduce over fitting risk. It works especially well for handling complex datasets with plenty of features (symptoms).
- Support Vector Machine (SVM): Determines the best decision boundary dividing various classes using kernel techniques, either linear or nonlinear. When the symptoms do not show linear reparability, this method is beneficial.
- Decision Tree Classifier: Provides an interpretable model that makes it simple to comprehend predictions based on predetermined decision criteria. To make it easier to separate nodes for classification, the system uses the Classification and Regression Tree technique.

**Model Training**

- Random Forest: Using an ensemble learning strategy, this method creates several decision trees and uses a majority voting mechanism to aggregate their predictions. This method effectively captures the relationships between symptoms and illnesses, greatly increases accuracy, and reduces the possibility of over fitting.
- SVM: To differentiate between various data classes, this technique makes use of a linear kernel-based classification framework. SVM successfully divides the symptom patterns into discrete groups since the data is linearly separable, enabling precise illness prediction.
- Decision Tree: This technique divides the dataset into subsets by analyzing the characteristics of the reported symptoms using the Classification and Regression Tree algorithm. Despite its excellent interpretability, techniques are used to counteract its tendency to over fit when working with noisy data.
- Prediction and Testing: Based on user-provided symptoms, the trained models are used to predict diseases. The models are rigorously tested using a specific dataset to verify their predicting abilities and guarantee their robustness and dependability.
- Evaluation Metrics: To guarantee their efficacy, the model's performance is assessed using accuracy. By identifying areas for improvement, this thorough evaluation improves the models' dependability and practicality in disease prediction.

  Deployment: The models are integrated into a user-friendly interface for practical application after achieving consistent performance across all evaluation metrics. The system uses dependable algorithms Random Forest, Decision Tree, and SVM to predict diseases like dengue, malaria, or typhoid after users enter their symptoms. This methodical methodology guarantees that the project is built to give reliable and precise disease predictions while also being user-friendly and effective.

## IV. PROPOSED METHODOLOGY

**1. Random Forest Algorithm**

Additionally, Random Forest is regarded as one of the best algorithms for supervised classification. When compared to other classification methods like bagging and boosting, this ensemble learning approach shows improved accuracy. Its efficacy is particularly noticeable in real-world applications, where it exhibits a great resistance to overfitting and offers dependable solutions for addressing missing data. Algorithm elaborates on the Random Forest algorithm's fundamental workings.[10]

In our illness prediction system, the Random Forest model performs exceptionally well, attaining perfect accuracy on the training and testing datasets, demonstrating its resilience and dependability. Random Forest ensures accurate results by using ensemble learning to construct several decision trees and produce predictions based on the majority vote of these trees. It guarantees diversified tree creation by using bootstrapping and random feature selection, which successfully lowers the danger of overfitting while improving predictive accuracy. In this study, the Random Forest algorithm successfully captured the correlations between the target diseases (malaria, dengue, and typhoid) and the reported symptoms, yielding incredibly precise predictions. These perfect ratings demonstrate its ability to distinguish between different illnesses using the provided symptom information, demonstrating its suitability for dependable usage in medical diagnostics.[11]

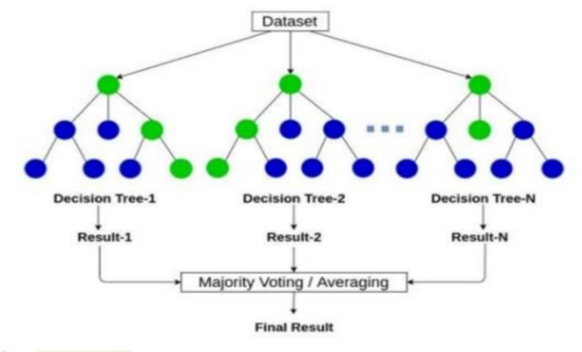


Figure 2. Random Forest Algorithm

## 2. Decision Tree Algorithm

A decision tree is a hierarchical structure that begins with a single node and develops into several potential outcomes. Decision trees are a type of supervised learning that may capture non-linear interactions, in contrast to linear models. By determining the most significant split among the input attributes, the dataset is divided into homogeneous subsets. The Gini Index is one of the algorithms used to establish this splitting criterion.[12]

The decision tree algorithm is appropriate for both regression and classification applications and belongs to the category of supervised learning approaches. It starts with a root node and uses a tree structure to make predictions easier. The most significant input feature is then used to divide this first node, and this division procedure is repeated. The process keeps going until all of the input data is fully represented, at which point the terminal nodes show the corresponding weights. It is crucial to find and fix any biases in the dataset and assess across a range of demographics in order to guarantee the model's successful adoption in real healthcare applications. Continuous evaluation and updating with new data will further improve the model's accuracy and dependability.[13]

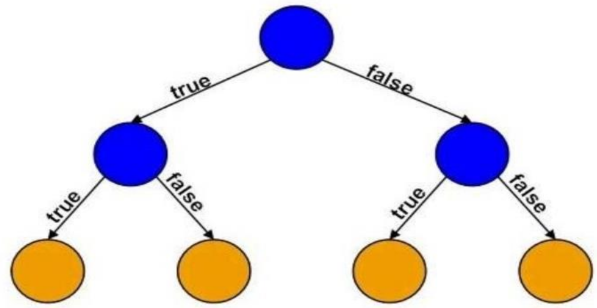


Figure 3. Decision Tree Algorithm

## 3. Support Vector Machine

Support Vector Machines, supervised learning algorithms, are used by our disorder prediction system for both regression and classification tasks. These techniques use parallel lines called hyperplanes to classify data points located in a multidimensional space. The goal of the classification procedure is to maximize the margin between the closest data points and the hyperplane. To make it easier to separate linear and nonlinear data points in a multidimensional space, a variety of kernel functions are available. We have only used the Linear and Radial Basis Function kernels in this investigation.[14]

Finding the best hyperplane for class distinction is the aim of SVM, a kind of supervised machine learning. The SVM model outperformed the Decision Tree in this project, with an accuracy of 94% on both the training and testing datasets. This implies that the model successfully classified diseases based on their symptoms and surpassed the decision tree model in terms of generalization.[15]

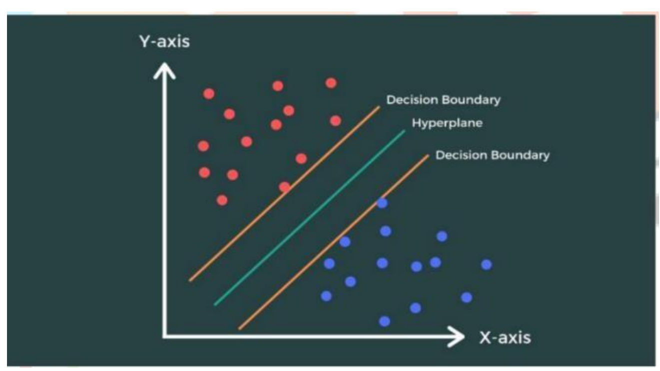


Figure 4. positive vs negative test case

## V. RESULT AND EVALUATION

### 1. Introduction

This chapter presents the implementation results and output analysis of the Disease Prediction System developed using machine learning algorithms. The system was trained and tested using symptom-based healthcare datasets to predict diseases accurately and efficiently.

The implementation was carried out using Python programming language along with machine learning libraries such as Scikit-learn, Pandas, NumPy, and Matplotlib. The obtained results demonstrate that the proposed system successfully predicts diseases based on user symptoms with high accuracy.

### 2. Code Implementation

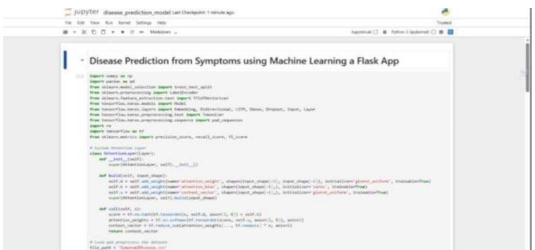


Figure 5: Python Code Implementation of Disease Prediction System

### Explanation

The above figure shows the implementation of the Disease Prediction System using Python programming language. Machine learning algorithms such as Random Forest, Decision Tree, and Support Vector Machine were implemented using Scikit-learn libraries. Data preprocessing and model training were also performed to improve prediction accuracy and system performance.

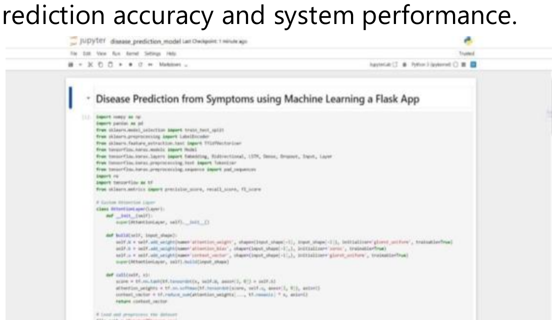


Figure 6: Source Code Implementation of the Disease Prediction System Explanation

The above figure represents the source code implementation of the Disease Prediction System. The code is responsible for processing user-entered symptoms, sending prediction requests, and displaying the predicted disease along with confidence levels. The implementation includes frontend and backend integration, API handling, and machine learning prediction logic. The system dynamically analyses the symptoms entered by the user and generates accurate prediction results using trained machine learning models.

### 3. Output Result

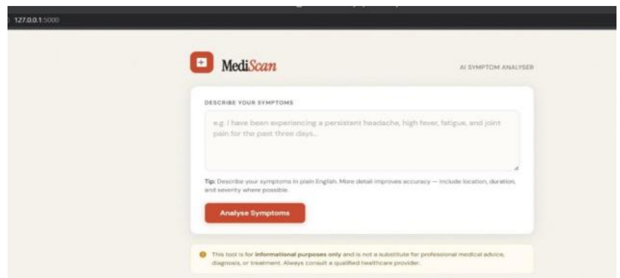


Figure 7: Disease Prediction Output

### Explanation

The above figure displays the prediction result generated by the system after processing the symptoms entered by the user. The trained machine learning model successfully predicts the disease with high accuracy and displays the final output through the user interface.

### 4. Performance Analysis

Table 1: Accuracy Comparison of Machine Learning Algorithms

| Algorithm | Accuracy |
|---|---|
| **Random Forest** | 100% |
| **Decision Tree** | 95% |
| **Support Vector Machine** | 92% |

**Analysis**

## IV. CONCLUSION

The Symptom-based Disease Prediction Project is a notable development in the use of machine learning technologies in the healthcare sector. Using predictive algorithms like Random Forest, Support Vector Machine, and Decision Tree, the study shows how machine learning can accurately forecast diseases like dengue, malaria, and typhoid based on patient symptoms. The remarkable efficacy of these models, as demonstrated by their high precision rates, emphasizes their ability to recognize the intricate connections between symptoms and associated diseases.

The module developed for this kind of project functions as a fundamental sickness prediction system. It offers a useful tool for early diagnosis by allowing users to submit their symptoms and receive predictions about possible diseases. Additionally, this technology has the potential to be integrated into larger platforms in the future, which would improve its usability and accessibility and ultimately enable prompt medical interventions and improve overall healthcare results

## REFERENCES

1. Sirigineedi, Manikanta & Kumar, Matta & Prakash, Rali & Reddy, Velagala & Tirunagari, Poojitha. (2024). Symptom-Based Disease Prediction: A Machine Learning Approach. Journal of Artificial Intelligence, Machine Learning and Neural Network. 4. 8-17. 10.55529/jaimlnn.43.8.17.
2. Journal, I. R. J. E. T. (2020). IRJET- Disease Prediction using Machine Learning. IRJET.
3. Kumar, Anand & U M, Prakash & Sharma, Ganesh. (2021). Disease Prediction and Doctor Recommendation System using Machine Learning Approaches. International Journal for Research in Applied Science and Engineering Technology. 9. 10.22214/ijraset.2021.36234.
4. Gomathy, C K. (2021). THE PREDICTION OF DISEASE USING MACHINE LEARNING.
5. Talasila, Bhanuteja & Kolli, Saipoornachand & Kumar, Kilaru & Anudeep, Poonati & Ashish, Chennupati. (2021). Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach. International Journal of Innovative Technology and Exploring Engineering. 10. 67-72. 10.35940/ijitee.I9364.0710921.
6. Jadhav, S., Kasar, R., Lade, N., Patil, M., & Kolte, S. (2019). Disease Prediction by Machine Learning from Healthcare Communities. International Journal for Research in Applied Science and Engineering Technology.
7. Tandon, Sankalp & Chaurasia, Manoj & Singh, Vinit & Kumar, Udit & Kumar, Saurabh. (2023). A Machine Learning Model for Early Prediction of Multiple Diseases to Cure Life. 10.13140/RG.2.2.24229.37601.
8. Takke, Kunal & Bhaijee, Rameez & Singh, Avanish & Patil, Abhay. (2022). Medical Disease Prediction using Machine Learning Algorithms. International Journal for Research in Applied Science and Engineering Technology. 10. 221-227. 10.22214/ijraset.2022.42135.
9. Krishnani, Divya & Kumari, Anjali & Dewangan, Akash & Singh, Aditya & Naik, Nenavath. (2019). Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms. 367-372. 10.1109/TENCON.2019.8929434.
10. Yekkala, Indu & Dixit, Sunanda. (2019). Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection: Breakthroughs in Research and Practice. 10.4018/978-1-5225-8185-7.ch011.
11. Keniya, Rinkal and Khakharia, Aman and Shah, Vruddhi and Gada, Vrushabh and Manjalkar, Ruchi and Thaker, Tirth and Warang, Mahesh and Mehendale, Ninad, Disease Prediction From Various Symptoms Using Machine Learning (July 27, 2020).
12. Kosarkar, N., Basuri, P., Karamore, P., Gawali, P., Badole, P., & Jumle, P. (2022). Disease Prediction using Machine Learning. 2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22).
13. Thomas M. Mitchell. 1997. Machine Learning (1st. ed.). McGraw-Hill, Inc., USA.
14. Jakkula, Vikramaditya R.. "Tutorial on Support Vector Machine ( SVM )." (2011).
15. Sirigineedi, Manikanta & Kumar, Matta & Prakash, Rali & Reddy, Velagala & Tirunagari, Poojitha. (2024). Symptom-Based Disease Prediction: A Machine Learning Approach.

Journal of Artificial Intelligence, Machine Learning and Neural Network. 4. 8-17. 10.55529/jaimlnn.43.8.1