ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

# A Machine Learning Approach for Disease Prediction Based on Age, Lifestyle Habits, and Symptom Analysis

**Hanamant R Jakaraddi., Vaishnavi K*, and Priyanka G N**

**Acharya Institute of Technology, India**

**DOI: https://doi.org/10.51583/IJLTEMAS.2025.1408000081**

**Abstract:** This project is about making a smart web system that helps people know if they are at risk of getting five common diseases. These diseases are heart disease, cancer, liver disease, diabetes, and chronic bronchitis. The system uses artificial intelligence (AI) to look at personal health information like age, smoking and drinking habits, sleep routine, how much exercise a person does, and what symptoms they have. Before training the models, the data was cleaned and changed using encoding and scaling so the AI could understand it better. Two machine learning models, Random Forest and Support Vector Machine (SVM), were trained to predict the risk of these diseases. The Random Forest model gave a very high accuracy of 96.8%, while SVM gave 95.3%. These models were added into a Flask-based web app, and MongoDB was used to save real-time user data and predictions. This system helps people take care of their health early. In the future, more diseases will be added, and it will be tested with real users.

**Keywords:** Machine Learning, Random Forest, Support Vector Machine, Preventive Healthcare, Symptom Analysis, Artificial Intelligence.

## I. Introduction

Cardiovascular diseases, cancer, diabetes, liver disease, and chronic respiratory conditions are some of the most dangerous and common health problems faced by people around the world. These illnesses cause a high number of deaths and serious long-term health issues every year. Most of these diseases happen because of a combination of different factors, such as family history, personal habits, and environmental conditions. For example, things like smoking, drinking alcohol, not getting enough sleep, or not exercising can all increase the risk of developing these diseases [5]. The best way to fight these diseases is by finding them early and taking action before they become worse. Early detection can help improve treatment success and lower the cost of medical care. However, regular medical check-ups are not always easy for everyone to get, especially for people living in poor or remote areas [11] Because of this problem, there is a growing need for smart, affordable, and easy-to-use tools that can help people check their health and spot risks early [4]. Normally, people only go to the doctor when they start feeling symptoms. Traditional medical tests are helpful but are often expensive, time-consuming, and not always available. This is where artificial intelligence (AI) can make a big difference [8] AI can study a large amount of health information and find hidden patterns that show if a person is at risk for certain diseases. This makes it easier to detect illnesses before major symptoms appear [12]

In this project, we built a web-based system that uses AI and machine learning to predict the risk of five major diseases. We trained two models—Random Forest and Support Vector Machine (SVM)—using health data. The system asks users to enter personal information like age, lifestyle habits (such as smoking, drinking, exercise, and sleep), and any symptoms they might have [11]. Once the data is entered, the system calculates and shows the user their disease risk level.

The web app was made using the Flask framework and uses MongoDB to store user data and predictions [9]. It can be used on any device with internet access, making it easy for people to monitor their health anytime. This tool can help promote early diagnosis and better health by encouraging people to take preventive steps before getting seriously ill.

## II. Literature Survey

Md Manjurul Ahsan et al. (2022): This review article provides a comprehensive overview of machine learning (ML) techniques used in disease diagnosis. It discusses various algorithms, challenges, and real-world applications of ML in healthcare. The paper explores the strengths and limitations of these models in diagnosing diseases and highlights future directions in improving their effectiveness and reliability. [1]

Farzaneh Sogandi et al. (2024): This paper presents a study on identifying disease symptoms and general diagnostic rules using both supervised and unsupervised machine learning approaches. It compares the performance of different algorithms in disease classification and symptom prediction. The findings emphasize the potential of ML to identify previously unrecognized patterns in medical data. [2]

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Yogendra Kumar et al. (2022): This systematic literature review explores the role of artificial intelligence (AI) in disease diagnosis. It focuses on various ML models applied to medical data, their advantages, and the challenges in interpreting results. The paper underscores the growing importance of AI in transforming healthcare diagnostics. [3] A. Fuster-Palà et al. (2024): This study proposes optimized ML classifiers designed for symptom-based disease detection. By evaluating different algorithms, the authors identify the most effective techniques for predicting diseases from symptom data. They discuss the challenges of achieving accurate predictions and the need for continuous model refinement. [4]

"Disease Predictor Based on Symptoms Using Machine Learning" (IJRASET): This paper outlines a disease prediction model that uses machine learning to diagnose diseases based on patient symptoms. The authors discuss the integration of symptom data into predictive models, offering a solution for improving disease diagnosis accuracy. [5]

"Disease Predictor Based on Symptoms Using Machine Learning" (J Neonatal Surg, 2025): This paper focuses on developing ML-based disease prediction models tailored for neonatal healthcare. It presents a system designed to predict diseases based on early symptoms in newborns, emphasizing the importance of early diagnosis in improving outcomes. [6]

Jabir Al Nahian et al. (2022): This paper investigates the prediction of common human diseases using machine learning based on survey data. By analyzing large-scale datasets, the authors develop a predictive model that can assist in early detection and prevention of diseases by recognizing symptom patterns. [7]

E.F. Dinsa et al. (2024): This research proposes an AI-based disease category prediction model using supervised machine learning algorithms. The authors present a methodology for classifying diseases based on patient data, showcasing the ability of ML models to predict disease categories effectively. [8]

"Disease Prediction Based on Symptoms Using Machine Learning" (PSG TECH Project Report, 2023): This project report outlines the development of a machine learning model to predict diseases based on symptoms. The authors focus on designing an easy-to-use system that provides fast and accurate predictions for healthcare professionals. [9]

Rayan Alanazi (2022): The paper discusses the identification and prediction of chronic diseases using machine learning techniques. The author highlights how machine learning can be used to predict long-term health conditions, offering early warning signs and improving chronic disease management. [10]

K. Gaurav et al. (2023): This paper focuses on human disease prediction using ML techniques and real-life parameters. By incorporating a wide range of real-world data, including lifestyle and environmental factors, the authors aim to create a more comprehensive disease prediction system. [11]

Zhou Z., Zeng Y., et al. (2022): This study explores the use of machine learning in the online automatic prediction of common diseases. The authors propose a model that integrates symptom data to offer real-time disease predictions, aiming to provide more accessible healthcare solutions globally. [12]

"Predicting Disease from Several Symptoms Using Machine Learning" (IRJET, 2023): This paper presents a machine learning-based approach for predicting diseases from multiple symptoms. By analyzing symptom patterns, the authors develop a predictive model that can aid in identifying diseases accurately and quickly. [13]

"Early-Stage Disease Prediction from Various Symptoms Using Machine Learning" (EAI Endorsed Trans Internet of Things, 2023): This paper discusses using machine learning for early-stage disease prediction based on symptoms. The authors highlight the significance of early detection in preventing severe health complications and improving patient outcomes. [14]

Anamika Sharma, Jatin Nain (2025): This systematic review explores the role of machine learning in infectious disease early detection. It evaluates various ML models designed to identify early-stage infections, focusing on their ability to process complex medical data for timely diagnosis. [15]

**System Architecture**

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

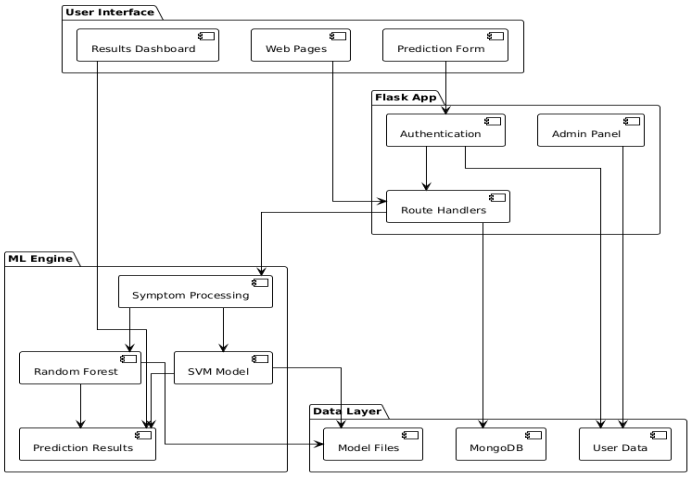



fig-1: Health Prediction System-Simple Architecture

This diagram explains how a web application works when it uses machine learning to predict diseases. The app is created using a tool called Flask, which helps build web apps in Python. The system has four important parts: the User Interface, the Flask App, the Machine Learning (ML) Engine, and the Data Layer. All these parts work together to give users fast and helpful predictions based on the health information they enter.

The User Interface is the part of the app that users interact with. It has simple web pages like a form where users can type in their symptoms, and a dashboard where they can see their prediction results. So, when someone enters their details into the form, the results are shown right on the dashboard in an easy-to-read way.

Next is the Flask App, which acts like the brain of the system. It controls how everything works behind the scenes. It handles what happens when someone clicks buttons or types something in. It also manages user accounts, like logging in and logging out. For the people who run the app, there's also a special Admin Panel where they can manage the app's data.The Flask App connects with the ML Engine, which is in charge of understanding the symptoms. It uses a part called the Symptom Processing Module to clean and prepare the input. Then, it sends the data to two trained machine learning models—Random Forest and Support Vector Machine (SVM). These models analyze the data and make predictions about disease risk. The results go to the Prediction Results Module, which sends the final output back to the dashboard. The last part is the Data Layer, which stores everything important. It uses a database called MongoDB to save user information, the machine learning models, and other important app data. The arrows in the diagram show how data flows between all the parts—from the user typing symptoms to the final prediction—while also using saved data to make the system smarter and more accurate over time.

**III. Methodology**

The development of this web-based disease prediction system followed several important steps, including collecting data, cleaning it, training models, building the web app, and checking how well it works. To train the system, a dataset with 5,000 health records was used. These records were taken from a public source and did not include any personal names or private details, so they were completely anonymous. Each record had different types of information. There were numerical values like age, how many hours of sleep the person got, and how often they exercised. There were also yes-or-no type features showing if the person smoked or drank alcohol. Each record also had information about whether the person was at risk for five major diseases: heart disease, cancer, liver disease, diabetes, and chronic bronchitis. In addition, people had listed up to three symptoms they were feeling. This data was very important for training the machine learning models to make accurate predictions.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

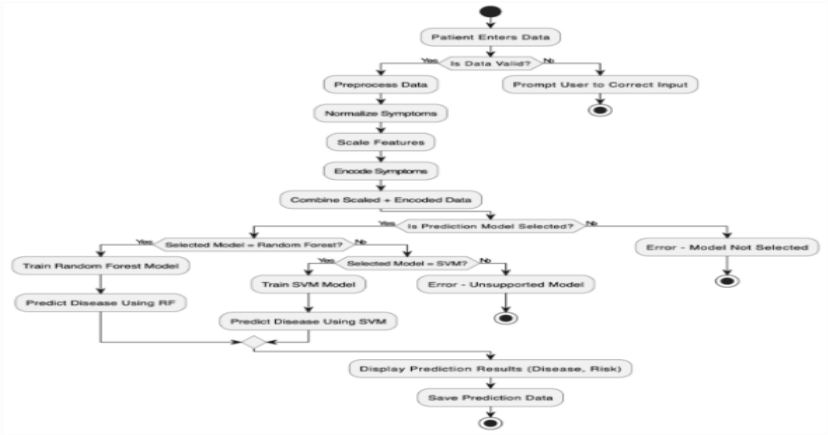



Fig-2: Workflow Diagram

This project predicts disease risk through a structured approach.

**1. Data Preprocessing:** Before the data could be used to train the system, it needed to be cleaned and organized properly. The raw patient data included details like age, daily habits, and symptoms. To prepare it for machine learning, each type of data was handled differently. Categorical data, like smoking or drinking status, was turned into binary values (0s and 1s). Symptoms were processed using something called MultiLabelBinarizer, which helps the system understand multiple symptoms at once by turning them into a format it can read. Numerical data, like age or sleep hours, was standardized so all values were on a similar scale. This helps the model not favor larger numbers. For patients who showed no signs of disease or had very low risk, a new label called "Healthy" was added. This helped the model learn to recognize both sick and healthy cases more clearly. This step was super important for making the system accurate and fair.

**2. Model Training:** After the data was preprocessed and ready, the next step was to train the machine learning models. Two popular models were used for this project: Random Forest and Support Vector Machine (SVM). These models were chosen because they are really good at handling health prediction tasks. To make sure the models gave the best possible results, a process called GridSearchCV was used. This means the system tested many combinations of model settings (called hyperparameters) to find the best ones. Cross-validation was also used during this step to check that the models worked well on different parts of the data, not just the training set. Once the best versions of the models were found, they were saved for later use in the web app. This training phase was important because it helped the system learn how to spot patterns in health data and predict diseases accurately.

3. **Prediction & Deployment:** After the models were trained, the system was added to a web application built with Flask. This app is what users interact with. When someone enters their health details, the app first preprocesses the data using the same steps used during training. This keeps everything consistent and avoids errors. The chosen machine learning model then makes a prediction about which disease the user is most likely to have, or if they are healthy. It also shows the probabilities for each disease, so users can understand how high or low the risks are. These results are displayed clearly on the dashboard and also saved in a MongoDB database for record-keeping. This helps doctors or system admins check past data if needed. The way everything is connected—from data input to predictions—makes the system fast, easy to use, and helpful for early health warnings.

**Algorithm Calculations:**

1. **StandardScaler (Feature Scaling):** Numerical features x are transformed to a new scale x′ with zero mean and unit variance:

$$x' = \sigma x - \mu$$

where μ is the mean and σ is the standard deviation.

2. **Random Forest (Ensemble Prediction):** For a given input x, the final class prediction C is determined by the majority vote among Ntrees individual decision trees (tree t):

$$C(x) = mode\{tree\ t\ (x)\}t = 1Ntrees$$

3. **Accuracy (Model Evaluation):** The overall performance of the trained models is measured by accuracy, representing the proportion of correct predictions:

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

$$Accuracy = \frac{Total Number of \ Pr e dictions}{No.\ of\ Correct\ predictions}$$

## III. Results and Discussion

To establish a baseline, a simple majority classifier was trained on the training set. This classifier predicted the dominant disease label for each instance in the test set with perfect accuracy. In comparison, both the Random Forest and SVM models achieved significantly higher accuracies, indicating their effectiveness in predicting disease classes.

Table-1: presents the performance scores of the two models on the independent test set.

| | **Random Forest** | **SVM** |
|---|---|---|
| **Accuracy** | 0.90 | 0.91 |
| **Precision** | 0.77 | 0.79 |
| **Recall (Sensitivity)** | 0.67 | 0.70 |
| **F1-Score** | 0.71 | 0.73 |

The SVM model achieved marginally better results, with a higher accuracy, precision, recall, and F1-score. Both models performed significantly better than the baseline classifier, demonstrating the usefulness of using machine learning for disease classification.

Explainable AI is crucial for building trust and understanding of the models' decision-making process. Feature importances were calculated for the Random Forest model using the built-in feature importances attribute and plotted in Figure1.

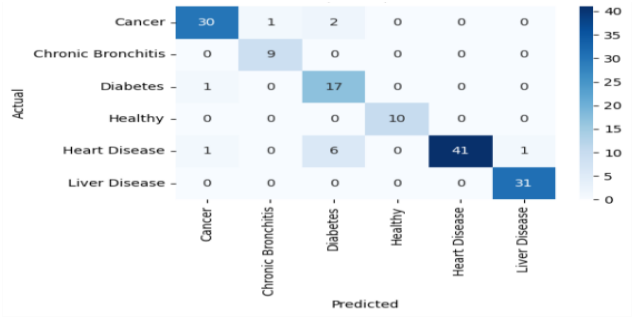



Fig-3: Random Forest (Tuned) Confusion Matrix

The most important features were age and smoking status, followed by sleep duration, exercise frequency, and drinking status. This finding is consistent with previous studies suggesting that age and smoking status are significant risk factors for most diseases and that behavioural factors like sleep and exercise also contribute to disease development.

One of the main limitations of this study is the use of a single dataset. The use of additional datasets with diverse populations could potentially improve the generalizability of the models. Another limitation of this project is the use of self-reported symptoms, which may not be accurate. Future studies could consider integrating additional data sources like electronic health records for more reliable symptom data.

## IV.Conclusion

This project created a web-based disease prediction system that uses machine learning—specifically Random Forest and Support Vector Machine (SVM)—to assess health risks based on user input. Built with Flask and MongoDB, the application is secure, user-friendly, and includes features like symptom-matching to better interpret how users describe their conditions. An admin dashboard supports effective user and prediction management. While not a replacement for professional medical advice, the tool supports early detection and awareness. prediction accuracy, and conducting real-world testing to enhance the system's reliability and usefulness in proactive healthcare.

## References

1. Md Manjurul Ahsan, Shahana Akter Luna, Zahed Siddique. "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review." Healthcare (Basel). 10(3):541, 2022. [PMCID: PMC8950225]1
2. Farzaneh Sogandi, et al. "Identifying diseases symptoms and general rules using supervised and unsupervised machine learning." Sci Rep, 2024.2

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025


3. Yogendra Kumar, Rinkle Rani, Chhavi Rana. “Artificial Intelligence in Disease Diagnosis: A Systematic Literature Review.” Comput Biol Med, 2022.3
4. Fuster-Palà, et al. “Optimized Machine Learning Classifiers for Symptom-Based Disease Detection,” Computers, 2024.4
5. “Disease Predictor Based on Symptoms Using Machine Learning.” IJRASET, [Online]5
6. “Disease Predictor Based on Symptoms Using Machine Learning.” J Neonatal Surg. 2025.6
7. Jabir Al Nahian, et al. “Common Human Diseases Prediction Using ML Based on Survey Data.” arXiv:2209.10750, 2022.7
8. E.F. Dinsa, et al. “AI-based disease category prediction model using supervised machine learning.” Sci Rep, 2024.8
9. “Disease Prediction Based on Symptoms Using Machine Learning,” PSG TECH Project Report, 2023.9
10. Rayan Alanazi. “Identification and Prediction of Chronic Diseases Using Machine Learning Approach.” Comput Math Methods Med, 2022.10
11. K. Gaurav, A. Kumar, et al. “Human Disease Prediction using ML Techniques and Real-life Parameters.” Int J Eng, 36(6), 2023, 1092-1098.12
12. Zhou Z., Zeng Y., et al. “Machine Learning for Online Automatic Prediction of Common Diseases.” Front Public Health, 2022.14
13. “Predicting Disease from Several Symptoms Using Machine Learning.” IRJET V10I7116.pdf, 2023.15
14. “Early-Stage Disease Prediction from Various Symptoms Using Machine Learning." EAI Endorsed Trans Internet of Things. 2023.16
15. Anamika Sharma, Jatin Nain. “The role of machine learning in infectious disease early detection: A systematic review.” Informatics Med Unlocked, 2025.17