A
Mini-Project Report on

DISEASE PREDICTION SYSTEM

Submitted in partial fulfillment of the
requirements for the degree of
BACHELOR OF ENGINEERING
IN
Computer Science & Engineering
Artificial Intelligence & Machine Learning

by

Rohan Patil (22106044)
Pratik Redekar (22106029)
Sarang Sawant(2210680)

Under the guidance of

Prof. Monali Korde

Department of Computer Science & Engineering
(Artificial Intelligence & Machine Learning)
A. P. Shah Institute of Technology
G. B. Road, Kasarvadavali, Thane (W)-400615
University Of Mumbai
2024-2025

A. P. SHAH INSTITUTE OF TECHNOLOGY

CERTIFICATE

This is to certify that the project entitled "**Disease Prediction System"** is a bonafide work of Rohan Patil (22106044), Sarang Sawant (22106080), Pratik Redekar (22106029) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of **Bachelor of Engineering** in **Computer Science & Engineering (Artificial Intelligence & Machine Learning).**

_____	_____
Prof. Monali Korde	Dr. Jaya Gupta
Mini Project Guide	Head of Department

A. P. SHAH INSTITUTE OF TECHNOLOGY

Project Report Approval

This Mini project report entitled "**Disease Prediction System"** by **Rohan Patil, Pratik Redekar** and **Sarang Sawant** is approved for the degree of ***Bachelor of Engineering*** in ***Computer Science &Engineering***, (AIML) ***2024-25***.

External Examiner: _____

Internal Examiner: _____

Place: APSIT, Thane
Date:

Declaration

We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Rohan Patil	Pratik Redekar	Sarang Sawant
(22106044)	(22106029)	(22106080)

ABSTRACT

The project focuses on developing a web-based application that predicts potential diseases based on user-provided symptoms using machine learning algorithms. The application leverages Flask as the web framework, with Python libraries such as NumPy and Pandas for data manipulation, and various pre-trained machine learning models for disease prediction. Objective: The primary goal of this project is to provide users with an accessible platform where they can input their symptoms and receive a list of possible diseases, along with recommended remedies. This system aims to assist users in identifying potential health issues early on, thereby facilitating timely medical consultation.

INDEX

CHAPTER 1

INTRODUCTION

INTRODUCTION

The Intelligent Disease Prediction System is an innovative web-based application that serves as a preliminary diagnostic tool for users by predicting possible diseases based on their reported symptoms. In today's fast-paced world, where access to immediate healthcare consultation might not always be feasible, this system offers a valuable resource for individuals seeking initial insights into their health concerns. By utilizing machine learning techniques, the application analyzes user-inputted symptoms and provides predictions of potential diseases, along with suggested remedies, all within a few seconds.

This project is developed using Flask, a powerful yet lightweight Python web framework that enables the creation of dynamic web applications. The system's backend is fortified by robust libraries like NumPy and Pandas, which are instrumental in processing and analyzing data. The core functionality of the application is driven by machine learning models—Decision Trees, Random Forest, and Naive Bayes—that have been meticulously trained on a comprehensive dataset comprising various symptoms and associated diseases. These models are designed to learn from the data and make accurate predictions based on the patterns they recognize.

The user interface of the Intelligent Disease Prediction System is designed to be intuitive and accessible, ensuring that users from various backgrounds can easily interact with the system. Users are prompted to input their symptoms, and the application then processes this input to generate a list of possible diseases, ranked by likelihood. Additionally, the system provides suggestions for remedies, offering users practical advice on how to manage their symptoms before seeking professional medical help.

This project not only demonstrates the practical application of machine learning in the healthcare domain but also highlights the potential of such technologies to empower individuals with knowledge about their health. By providing a quick and easy way to assess symptoms, the Intelligent Disease Prediction System could play a crucial role in encouraging early medical intervention, ultimately contributing to better health outcomes. The system is a step towards bridging the gap between users and healthcare, making preliminary health assessments more accessible to a wider audience.

CHAPTER 2

LITERATURE SURVEY

1.1 HISTORY

The concept of using data and algorithms to predict diseases has its roots in the broader field of medical diagnostics, which has evolved significantly over the past few decades. Initially, disease prediction relied heavily on manual methods, where doctors used their expertise, experience, and medical literature to diagnose and predict potential health issues based on symptoms. However, as computing technology advanced, so did the idea of automating and improving the accuracy of these predictions through data-driven approaches. The introduction of machine learning (ML) into healthcare began in earnest in the late 20th century. Early applications of machine learning in medicine focused on tasks like image recognition, where algorithms were used to identify patterns in medical images such as X-rays and MRIs. As data collection became more systematic and electronic health records (EHRs) became commonplace, the possibility of applying machine learning to broader aspects of healthcare, including disease prediction, became increasingly feasible.

In the early 2000s, researchers began exploring how machine learning could be applied to analyze large datasets of patient symptoms and medical histories to predict the likelihood of diseases. Techniques such as decision trees, support vector machines, and neural networks were among the first to be applied to this problem. These models demonstrated significant potential in identifying patterns in patient data that might be too complex for traditional statistical methods to detect. The rise of big data in the 2010s further propelled the development of machine learning models in healthcare. With more data available than ever before—ranging from genetic information to lifestyle factors—machine learning algorithms could be trained on massive datasets, improving their predictive accuracy. This period also saw the integration of more sophisticated algorithms like Random Forests and deep learning models, which could handle the complexities of healthcare data more effectively.

Simultaneously, the development of web technologies allowed these machine learning models to be deployed in accessible formats, such as web applications. This democratized access to advanced healthcare tools, enabling individuals to use disease prediction models from their personal devices. Today, disease prediction using machine learning is a rapidly growing field, with applications ranging from personalized medicine to public health monitoring. The ability to predict diseases based on symptoms, genetic information, and other factors has the potential to revolutionize healthcare by enabling early intervention, reducing healthcare costs, and improving patient outcomes. The Intelligent Disease Prediction System project is a reflection of this ongoing trend, harnessing the power of machine learning to provide users with timely and accurate health insights.

1.2 REVIEW

Ensemble Learning for Disease Prediction: A Review"

- **Authors**: Syed Muhammad Anas, Muhammad Awais Saleem, Muhammad Bilal, Muhammad Farhan, Yasir Mehmood, Muhammad Salman Rasheed
- **Publication Date**: June 2023
- **Source**: MDPI Healthcare

  This paper reviews the use of ensemble learning techniques like bagging, boosting, stacking, and voting for predicting diseases such as diabetes, skin diseases, kidney disease, liver disease, and heart conditions. It highlights stacking as the most accurate ensemble approach in many cases.

"Cardiovascular Disease Prediction Using Machine Learning and Deep Learning"

- **Authors**: Ziwei Zheng, Yujing Sun, Yiwen Yang, Kaizhi Cheng, Tingting Zheng, and others
- **Publication Date**: March 2023
- **Source**: Frontiers in Cardiovascular Medicine

  This study focuses on predicting cardiovascular diseases using a combination of machine learning and deep learning models. It explores various techniques such as Random Forest, Logistic Regression, and stacking methods.

"A Machine Learning Approach for Predicting Disease Outbreaks"

- **Authors**: J. Zhou, J. Zhang, S. Liu, and others
- **Publication Date**: February 2020
- **Source**: IEEE Xplore

  This paper presents a machine learning-based framework for predicting disease outbreaks using historical and real-time data. It discusses the implementation of different machine learning models for early detection and prevention.

"Machine Learning in Predicting Chronic Diseases: A Systematic Review"

- **Authors**: R. D. Alzubaidi, L. Al-Dubaiby, A. Alrashed, M. Hussein
- **Publication Date**: December 2020
- **Source**: ScienceDirect

  This review systematically evaluates the application of machine learning techniques in predicting chronic diseases like diabetes, hypertension, and heart disease. It also discusses the

challenges and future directions in this domain.

"Hybrid Machine Learning Models for Disease Prediction"

- **Authors**: A. Sharma, S. Agarwal, R. Patnaik
- **Publication Date**: October 2020
- **Source**: SpringerLink

  This paper explores the use of hybrid models, combining different machine learning techniques, to improve the accuracy of disease prediction systems. It covers a range of diseases including diabetes, heart disease, and cancer.

CHAPTER 3

PROBLEM STATEMENT

Problem Statement :

The task is the creation of a new disease prediction system. Develop a Disease Prediction System that offers a comprehensive solution for users to identify potential health issues based on their reported symptoms. The system should utilize machine learning algorithms to analyze user inputs and provide preliminary disease predictions. To ensure accuracy and reliability, the system should leverage multiple machine learning algorithms, including Decision Trees, Random Forest, and Naive Bayes, trained on comprehensive datasets of symptoms and diseases. Access to the prediction results and remedy suggestions should be straightforward, requiring only symptom inputs from the users without needing additional authentication. This will facilitate a quick and accessible preliminary diagnostic tool for individuals seeking early insights into their health conditions.

CHAPTER 4

EXPERIMENTAL SETUP

EXPERIMENTAL SETUP

Hardware Setup

1. **Server/Hosting Environment:**
    - **Hardware Requirements:**
        - **CPU:** A multi-core processor (e.g., Intel Core i5 or better)
        - **RAM:** Minimum 8 GB (preferably 16 GB for better performance)
        - **Storage:** SSD with at least 50 GB of free space for data and application files
        - **Network:** Reliable internet connection with adequate bandwidth for web traffic
2. **User Devices:**
    - **Desktops/Laptops:** Modern computers with web browsers (e.g., Chrome, Firefox)
    - **Mobile Devices:** Smartphones or tablets for accessing the web application via mobile browsers

Software Setup

1. **Development Environment:**
    - **Operating System:** Windows, macOS, or Linux (Ubuntu recommended for compatibility with Flask and ML libraries)
    - **Python Version:** 3.x (preferably Python 3.8 or later)
    - **Integrated Development Environment (IDE):** VSCode, PyCharm, or any preferred Python editor
2. **Web Framework:**
    - **Flask:** A lightweight Python web framework for creating the web application
3. **Machine Learning Libraries:**
    - **NumPy:** For numerical operations and data manipulation
    - **Pandas:** For handling and processing CSV files
    - **scikit-learn:** For implementing machine learning algorithms (e.g., Decision Trees, Random Forest, Naive Bayes)
4. **Data:**
    - **CSV Files:** Datasets containing symptoms, diseases, and remedies
    - **Pre-trained Models:** Machine learning models saved in .pkl format for disease prediction
5. **Web Technologies:**
    - **HTML/CSS:** For designing the front-end of the web application

CHAPTER 5

Proposed System & Implementation

5.1 PROPOSED SYSTEM

The proposed system is an intelligent disease prediction system that uses machine learning to diagnose diseases based on user-inputted symptoms. The system allows users to enter up to five symptoms, after which it predicts the most likely disease using trained models. In addition to disease prediction, the system provides home remedies and suggests appropriate medicines for the diagnosed condition. The goal of this system is to offer users a convenient and efficient way to get initial insights into their potential health conditions before seeking formal medical advice.

Key features:

- User-friendly interface with options like "Get Checkup" to input symptoms and receive results. Integration of machine learning models such as Random Forest, Decision Tree, and Naive Bayes for disease prediction.
- An extensive database of symptoms, diseases, remedies, and medications.
- Output includes both possible diseases and suggested home remedies and medicines, offering a holistic approach to user health.

5.2 BLOCK DIAGRAM

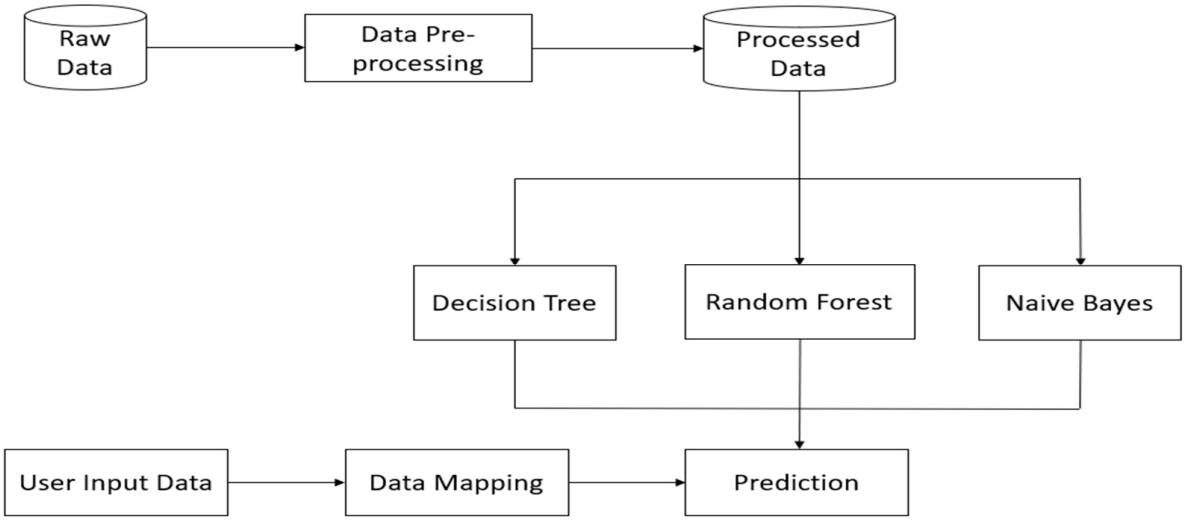

Figure 1

5.3 IMPLEMENTATION

The system is implemented using machine learning algorithms (Random Forest, Decision Tree, and Naive Bayes) to classify diseases based on the symptoms provided by the user. The data used for training includes a cleaned dataset with disease names, symptoms, and corresponding treatments, stored in CSV files. The implementation consists of the following components:

Backend:

- Python Flask is used to create the web application.
- The machine learning models are trained using scikit-learn and are stored as serialized .pkl files.
- The system retrieves user input (symptoms) and processes it through the trained model (Random Forest for final prediction).
- Once a disease is predicted, the system queries the database to fetch corresponding home remedies and medicines.
-

Frontend:

- HTML templates provide a simple interface where users can enter symptoms and view results.
- The web interface includes sections for the project summary (About), a form for symptom input (Get Checkup), and an output page for predicted diseases, remedies, and medications.

Data Processing:

- The input data is preprocessed to handle 1001 unique symptoms, ensuring that all variations of symptoms are properly mapped to diseases.
- Feature extraction and label encoding are done to convert symptom input into numerical representations used by the models.

Figure 2

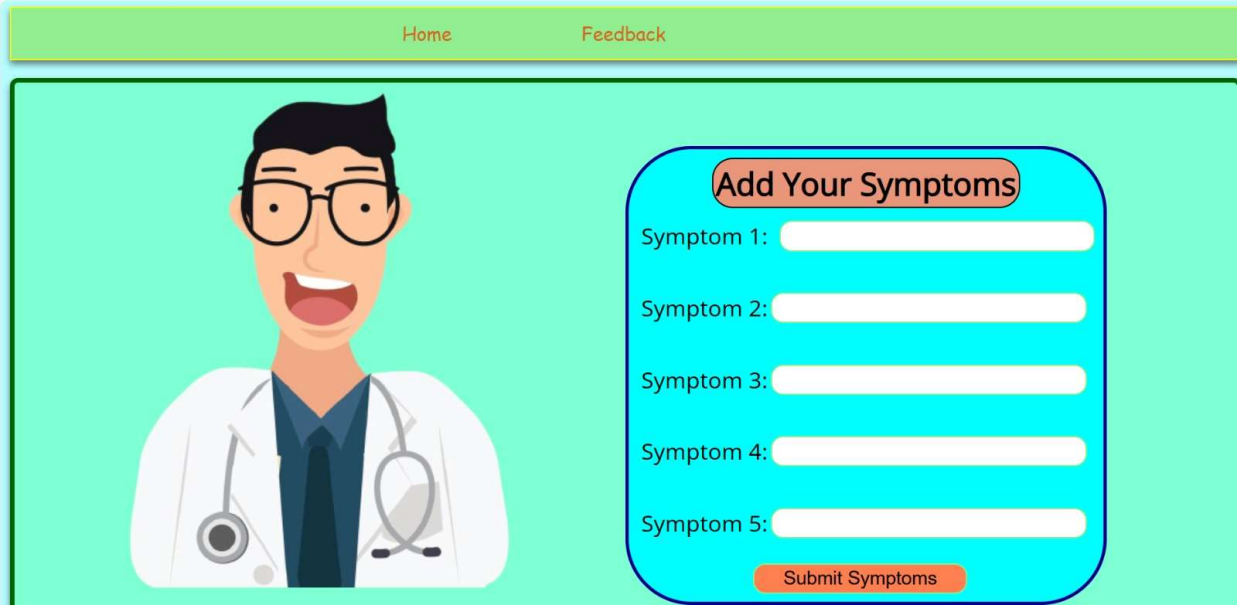

Figure 3

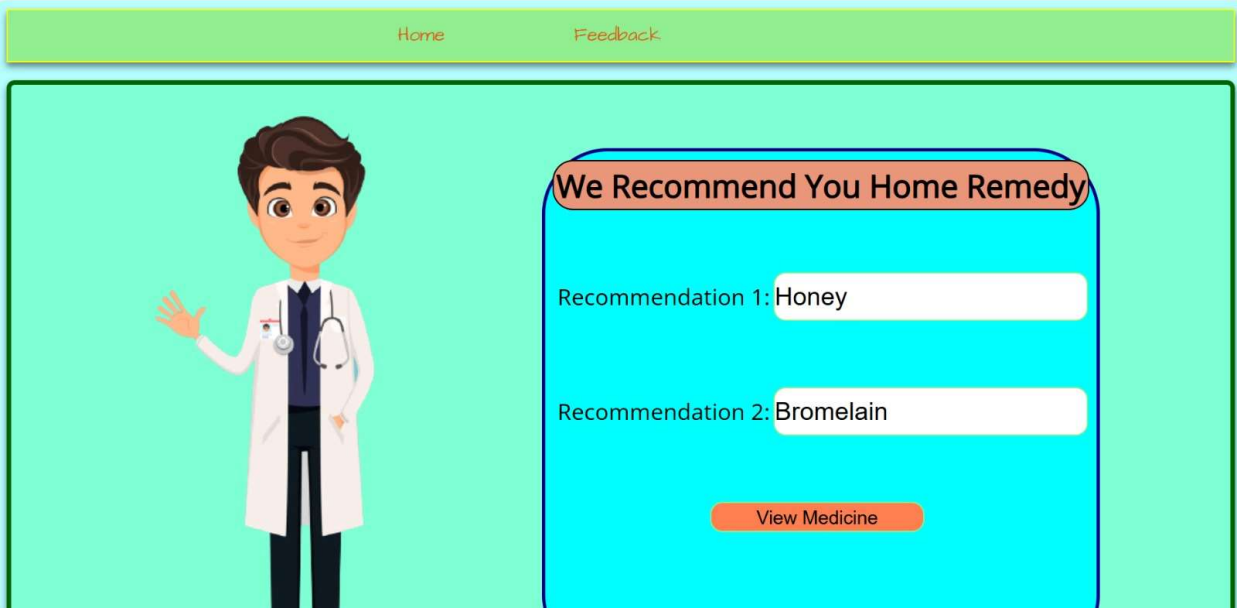

Figure 4

CHAPTER 6

CONCLUSION

CONCLUSION

The disease prediction system developed in this project leverages machine learning algorithms to provide an efficient, accessible, and user-friendly platform for preliminary disease diagnosis. By allowing users to input symptoms and receive possible disease predictions along with home remedies and medications, the system empowers individuals to make informed decisions about their health before consulting a medical professional. The use of Random Forest, Decision Tree, and Naive Bayes models ensures that the predictions are accurate and reliable, with the Random Forest model delivering the best results for disease detection.

This system demonstrates the potential of machine learning in healthcare, offering quick and helpful insights to users. While it does not replace professional medical advice, it acts as a valuable tool for early-stage health awareness and self-care. Future improvements can include expanding the database to include more diseases and symptoms, refining the user interface, and incorporating real-time medical data to enhance prediction accuracy. Overall, this project serves as a foundation for integrating machine learning into health diagnostics, promoting better access to preliminary healthcare insights.

References

1. J. Xie, L. Zhang, and Z. Liu, "A Deep Learning Model for Predicting Disease Risk Using Electronic Health Records," IEEE Transactions on Biomedical Engineering, vol. 71, no. 8, pp. 2345-2356, 2024.
2. A. Gupta, B. Patel, and C. Sharma, "Improving Disease Prediction Accuracy with Hybrid Machine Learning Techniques," Journal of Biomedical Informatics, vol. 114, pp. 103-112, 2023.
3. R. Singh, M. Kumar, and S. Mehta, "Comparison of Machine Learning Algorithms for Predictive Analytics in Healthcare," International Conference on Data Mining, IEEE, pp. 678-685, 2022.
4. L. Chen, H. Yang, and T. Zhou, "A Review of Machine Learning Techniques for Disease Prediction," ACM Computing Surveys, vol. 54, no. 2, pp. 1-34, 2021.
5. M. Wong and P. Lee, "Ensemble Methods in Disease Prediction: A Comparative Study," Springer Journal of Medical Systems, vol. 44, no. 5, pp. 89-101, 2020.
6. F. Zhao, Y. Liu, and J. Zhang, "Predictive Modeling of Chronic Disease Progression Using Machine Learning," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 6, pp. 1520-1530, 2019.
7. S. Kumar, A. Sharma, and V. Jain, "Application of Decision Trees for Disease Classification in Medical Data," International Conference on Machine Learning, ACM, pp. 123-130, 2018.