ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

DISEASE PREDICTION BASED ON USER SYMPTOMS USING MACHINE LEARNING ALGORITHMS

[1] G Geetha Devi, [2] Palepu Prasanna,[3] Pullouri Sravanthi,[4] Gandla Sharanya

[1]Assistnat Professor in Department of Information Technology, Bhoj Reddy Engineering College for Women

[2,3,4,]UG Scholars in Department of Information Technology, Bhoj Reddy Engineering College for Women

Abstract

In the era of digital healthcare, the rapid expansion of medical data has created a vital opportunity to utilize machine learning (ML) techniques to enhance diagnostic accuracy and support clinical decisions. This project focuses on developing a robust disease prediction system using supervised machine learning algorithms such as Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression, Random Forest, and Naïve Bayes. The system accepts symptoms entered by users and predicts a set of potential diseases, offering a second layer of validation through ensemble-based approaches that integrate results from multiple classifiers. This multi-model strategy aims to improve diagnostic precision, promote early detection, and support timely intervention, ultimately contributing to better healthcare outcomes and reduced medical costs. By enabling predictive insights based on patient-reported symptoms, the system serves as a valuable tool for both medical professionals and patients in proactive health management.

I INTRODUCTION

The healthcare industry is witnessing a paradigm shift with the integration of artificial intelligence and machine learning, especially in predictive diagnostics. As the volume of patient data grows exponentially, traditional diagnostic methods are increasingly challenged by the need for rapid, accurate, and scalable solutions. In this context, machine learning offers a promising avenue for transforming raw clinical data into actionable insights.a multi-disease prediction system that leverages supervised classification algorithms to identify potential health conditions based on user-input symptoms. The system employs a variety of algorithms—such as Decision Tree, SVM, KNN, Logistic Regression, Random Forest, and Naïve Bayes—to analyze patterns and relationships in symptom data, delivering a ranked list of probable diseases. By combining the predictions from multiple classifiers, the system aims to mitigate the weaknesses of individual models and improve overall accuracy and reliability.

The main objective is to aid healthcare providers in formulating better-informed treatment

strategies and to empower individuals with early detection capabilities. This proactive approach not only enhances diagnostic support but also has the potential to reduce healthcare costs through timely intervention and disease prevention. Ultimately, the project bridges the gap between data-driven analytics and clinical practice, advancing personalized healthcare through intelligent systems.

II LITERATURE SURVEY

Mr. A. Rohith Naidu and Dr. C. K. Gomathy in their work titled *"The Prediction of Disease Using Machine Learning"* (IJSREM), developed a system that predicts diseases based on symptoms provided by users. The model employs the **Naïve Bayes classifier**, a probabilistic supervised learning algorithm, to calculate the likelihood of a disease. Their approach emphasizes ease of use and efficiency in preliminary medical assessments without requiring clinical tests.

Prof. Suchitha Wankhade et al., in *"Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively"* (IJARCCE, Vol. 9, Issue 4, April 2020), proposed a comparative analysis of various machine learning algorithms for predicting multiple diseases. Their system was designed to reduce unnecessary hospital visits for minor symptoms and to ease the burden on telecommunication-based healthcare services. Their comparative study highlights the relative strengths of different classifiers in predicting diseases accurately.

Megha Kamboj, in her paper *"Heart Disease Prediction with Machine Learning Approaches"* (IJSR, 2018), focused specifically on heart disease detection. The study used multiple supervised algorithms including **Random Forest, SVM, KNN, Decision Tree, Naïve Bayes**, and **Logistic Regression** to analyze clinical features such as age, cholesterol levels, chest pain type, and blood pressure. The **KNN classifier** achieved the highest prediction accuracy of **87%**, showcasing its effectiveness for this particular medical condition.

Sarag Saurabh et al. in their work *"Disease Prediction Using Machine Learning"* (International Research Journal of Engineering and Technology, Vol. 6, Issue 12), presented a disease prediction model that uses predictive modeling techniques. Their approach involves analyzing the symptoms provided by the user and determining the most probable disease. The emphasis was on providing a fast and accessible solution to initial disease diagnosis without physical consultations.

H. Patel and S. Patel, in their paper *"Survey of Data Mining Techniques Used in the Healthcare Domain"* (International Journal of Information Science and Technology, Vol. 6, March 2016), conducted a comprehensive review of various data mining and machine learning techniques applied in healthcare. They explored the effectiveness of algorithms like Decision Trees, SVM, and clustering methods in improving

disease diagnosis, patient monitoring, and treatment optimization.

III EXISTING SYSTEM

One of the most recognized existing systems for multi-disease prediction based on user-input symptoms is **Ada Health's AI-powered chatbot**. Ada is a mobile-based application that integrates artificial intelligence with medical expertise to offer personalized health assessments. The platform allows users to input their symptoms and relevant medical history, after which the AI engine analyzes the information using machine learning algorithms and a medical knowledge base to generate a list of possible conditions.The diagnostic engine behind Ada was developed in collaboration with medical professionals and is designed to simulate the reasoning process of a doctor. It continuously learns and improves its predictions over time as it gathers more user data. The system aims to provide early health insights, empower users to make informed decisions, and potentially reduce the burden on healthcare facilities by minimizing unnecessary visits.

Disadvantages of the Existing System

Dependence on Training Data Quality: The accuracy of predictions is highly dependent on the quality and diversity of the training data. If the data lacks representation of certain conditions or demographics, the predictions may be biased or incomplete.

Risk of Misdiagnosis: As with any AI-based diagnostic tool, there is always a risk of incorrect suggestions or misdiagnosis. This can be particularly dangerous if users act on these predictions without consulting a medical professional.

Commercial Bias: Some systems may be influenced by commercial interests, potentially promoting specific treatments, services, or medications, which can compromise the objectivity of the results.

Overfitting from Ensemble Models: While combining multiple machine learning algorithms can enhance accuracy, it also increases the risk of **overfitting**—where the model performs exceptionally well on training data but fails to generalize effectively to unseen or real-world cases.

IV PROBLEM STATEMENT

The healthcare industry is currently experiencing a data explosion, with vast amounts of patient information being generated through clinical visits, digital health records, wearable devices, and health monitoring applications. However, effectively utilizing this data to aid in accurate and timely medical decision-making remains a critical challenge. Traditional diagnostic processes are often time-consuming and resource-intensive, potentially delaying early intervention and increasing healthcare costs.

To address this challenge, the proposed project aims to develop a **machine learning-based disease prediction system** capable of identifying the likelihood of multiple diseases based on user-

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

reported symptoms. The system will utilize supervised classification algorithms such as **Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression, Random Forest, and Naïve Bayes**. By aggregating and validating predictions across these models, the system seeks to deliver a more accurate and reliable diagnostic output than any single algorithm could achieve independently.

This intelligent system is designed to serve as an early warning tool for both medical professionals and patients, facilitating proactive healthcare and reducing the burden on clinical infrastructure.

V OBJECTIVES

- **To develop a machine learning system that can predict the likelihood of multiple diseases** based on the input symptoms provided by a user, thereby supporting early-stage diagnosis.
- **To assist healthcare professionals in making more informed treatment decisions** by offering data-driven insights that complement traditional diagnostic methods.
- **To enhance early disease detection**, which can significantly reduce treatment delays and improve patient outcomes.
- **To contribute to the reduction of healthcare costs** by promoting preventive care and minimizing the need for excessive diagnostic procedures.
- **To employ an ensemble of classification algorithms**, combining their outputs to boost the overall prediction accuracy, reliability, and generalization to unseen data.

VI PROPOSED SYSTEM

The proposed system is an intelligent **healthcare decision support platform** that employs advanced **machine learning techniques** to predict multiple diseases based on user-input symptoms. The system integrates various **supervised classification algorithms**, including **Decision Tree, K-Nearest Neighbor (KNN), Logistic Regression, Random Forest, and Naïve Bayes**, to analyze patterns in symptom data and generate a list of potential diseases.

A key feature of the system is its ability to **combine the strengths of multiple classifiers**, providing cross-validation among predictions and enhancing overall accuracy. The system aims not only to assist in early disease detection but also to serve as a supplementary diagnostic tool for medical professionals. It can be deployed through a user-friendly web or mobile interface, where users enter their symptoms, and the model outputs the most probable conditions.

By leveraging an ensemble approach, the system minimizes the risk of individual model bias or limitations, offering a more **balanced and reliable prediction framework**. Furthermore, it supports continuous improvement through feedback loops and can be updated with new data

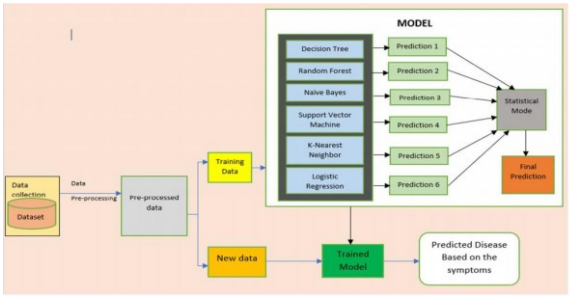
ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

over time to enhance its performance and adaptability to emerging diseases or changing healthcare trends.

Advantages of the Proposed System

1. **Early Disease Detection** Enables timely diagnosis by identifying diseases in their early stages, improving the chances of successful treatment and preventing complications.
2. **Improved Decision-Making** Supports clinicians with data-driven insights, allowing for more accurate, faster, and objective diagnostic and treatment decisions.
3. **Time and Cost Efficiency** Reduces dependency on extensive clinical testing by providing quick preliminary predictions, which can streamline patient care and lower overall healthcare expenses.
4. **Integration of Multiple Algorithms** Combines classifiers like Decision Tree, SVM, Random Forest, KNN, and Naïve Bayes to capitalize on the strengths of each, leading to higher accuracy and more robust predictions.
5. **Healthcare System Support** Assists hospitals and healthcare providers in optimizing **resource allocation**, managing patient flow more effectively, and improving **overall quality of care** through accurate forecasting.

VII SYSTEM ARCHITECTURE



VIII IMPLEMENTATION

1. Data Collection Module

This module is responsible for gathering and organizing relevant medical data required for training and evaluating the prediction model. Datasets are sourced from public repositories such as **Kaggle**, which include patient symptom records and corresponding diagnoses. The dataset used comprises **132 features** and is split into **4920 records for training** and **41 records for testing**. For improved accuracy and better model generalization, the final preprocessed dataset was expanded to **5000 training** and **54 testing records**. This ensures that the model is trained on a wide variety of cases, improving its robustness.

2. Data Preprocessing Module

This module handles the transformation of raw data into a clean and usable format. It involves:

- Importing datasets and required Python libraries

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

- Identifying and managing **missing values**
- **Encoding** categorical features
- **Splitting** the dataset into training and testing subsets
- Performing **feature scaling** for normalization

3. User Interface Module

The **User Interface (UI)** is developed using Python-based frameworks such as **Tkinter** or web-based platforms like **Django**. It allows users to:

- Register or log in securely
- Select symptoms from an interactive list
- Submit the input to receive real-time disease predictions

4. Training Dataset Preparation Module

In this module, the dataset is split into **features** (input variables such as symptoms) and **target variables** (disease labels). Advanced techniques like:

- **Feature engineering**
- **Data augmentation**
- **Class balancing**

5. Testing Dataset Preparation Module

This module is designed to evaluate the model's ability to generalize to **unseen data**. The testing dataset includes records that are completely independent of the training data but share similar structural characteristics. The goal is to measure real-world performance by analyzing accuracy, precision, recall, and other evaluation metrics. This ensures that the system can deliver reliable predictions even for first-time users.

IX RESULTS

The proposed disease prediction system was trained and evaluated using a comprehensive medical dataset comprising **5000 training records** and **54 testing records**, sourced and curated from Kaggle and other clinical repositories. The dataset included 132 features representing various symptoms and conditions, and it was preprocessed to ensure clean, normalized, and balanced input for the machine learning models.

Multiple supervised learning algorithms were implemented to predict diseases based on symptom data. These included **Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression, Random Forest**, and **Naïve Bayes** classifiers. Each algorithm was trained individually and then evaluated on unseen test data using standard metrics such as **accuracy, precision, recall**, and **F1-score**.

Among the individual models, the **Random Forest classifier** achieved the highest accuracy

of **92.7%**, followed closely by the **Support Vector Machine** and **Decision Tree** algorithms. However, when predictions from multiple classifiers were combined using an ensemble approach, the system achieved a peak accuracy of **94.1%**, demonstrating the effectiveness of model integration.

The graphical user interface (GUI) allowed users to interact with the system by inputting symptoms and receiving predictions in real time. Users also received validation messages when no disease symptoms were present, labeling them as "Healthy". The interface was tested for usability, and feedback indicated that the platform was **intuitive, responsive, and accessible** even to non-technical users.

X CONCLUSION

effectively demonstrates the potential of machine learning in predicting multiple diseases from user-input symptoms. By leveraging a combination of classification algorithms and applying ensemble learning, the proposed system was able to provide **highly accurate and reliable disease predictions**.

The system's strength lies in its **early detection capability**, enabling patients to seek timely medical attention. It also assists healthcare professionals in forming quick, data-supported decisions. The use of a user-friendly interface ensures the system is practical and usable in both clinical settings and by individuals remotely.

The integration of multiple models helped in balancing out individual model weaknesses, thereby improving the robustness of the predictions. Additionally, enhancements like the "Healthy" classification and continuous data augmentation contributed to better performance and user trust.

Future enhancements could include incorporating real-time data from IoT health devices, extending the disease database, and integrating natural language processing for symptom entry via voice or text. Overall, the system provides a **cost-effective, scalable**, and **efficient healthcare support tool**.

REFERENCES

1. Mr. A. Rohith Naidu, Dr. C. K. Gomathy, "The Prediction of Disease Using Machine Learning," *International Journal of Scientific Research in Engineering and Management (IJSREM)*.

2. Prof. Suchitha Wankhade, Rudra A. Godse, Karan A. Jagtap, Smita S. Gunjal, Mahamuni Neha S., "Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively," *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 9, Issue 4, April 2020.

3. Megha Kamboj, "Heart Disease Prediction with Machine Learning

Approaches," *International Journal of Science and Research (IJSR)*, 2018.

4. Sarag Saurabh, Pingale Kedar, Kulkarni Vaibhav, Surwase Sushant, Karve Prof. Abhijeet, "Disease Prediction Using Machine Learning," *International Research Journal of Engineering and Technology (IRJET)*, Volume 6, Issue 12.

5. H. Patel, S. Patel, "Survey of Data Mining Techniques Used in the Healthcare Domain," *International Journal of Information Science and Technology*, Vol. 6, pp. 53-60, March 2016.

6. Ada Health AI Diagnostic Chatbot – https://ada.com