ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Multiple Disease Prediction Using Machine Learning Techniques

Mr.K. Shiva Prasad[1], K.Ruchitha Devi[2], P.Meghana[3], Md.Basharath Hussain[4]

[1]Assistant Professor Department of CSE(AI &ML) Keshav Memorial Engineering College Osmania University, Hyderabad

[2,3,4]Department of CSE(AI &ML) Keshav Memorial Engineering College Osmania University, Hyderabad

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150500085

Received: 06 May 2026; Accepted: 11 May 2026; Published: 02 June 2026

ABSTRACT

The growing demand for early diagnosis and data-driven clinical support has positioned machine learning as a compelling tool in modern healthcare. Most existing disease prediction systems, however, are built around a single classification model that produces one deterministic output — an approach that inherently fails to reflect the uncertainty and symptomatic overlap commonly encountered in real-world clinical scenarios. This limitation restricts both patients and healthcare providers from meaningfully considering alternative probable conditions during preliminary assessment, and disproportionately affects individuals in underserved regions where timely access to professional medical consultation remains scarce. To address this gap, the proposed system employs a soft-voting ensemble framework that integrates three complementary classifiers — Decision Tree, Naive Bayes, and Random Forest — to generate more balanced and probabilistically informed predictions. Given symptom-based inputs, the system identifies and ranks the top four probable diseases, offering a broader diagnostic perspective than a single-model approach could provide. A real-time web interface allows users to select their symptoms and instantly view ranked predictions, supporting informed preliminary self-assessment and facilitating more meaningful consultations with healthcare professionals.

Keywords— Machine Learning, Disease Prediction, Soft Voting Ensemble, Healthcare Decision Support.

INTRODUCTION

Timely and accurate disease diagnosis remains one of the most critical factors in determining patient outcomes and preventing the escalation of severe medical complications. In conventional healthcare systems, disease identification relies primarily on clinical examinations, laboratory tests, and the expertise of trained medical professionals. While these methods are largely effective, they are often time-consuming, financially burdensome, and physically inaccessible to individuals living in remote or underserved regions. Consequently, many patients face significant delays in receiving adequate medical guidance, which can contribute to the progression of underlying conditions and place an increasing strain on already overburdened healthcare systems.

A further challenge inherent to conventional diagnosis is the considerable complexity and symptomatic overlap that exists across many diseases. Numerous illnesses share common early presentations such as fever, fatigue, headache, and nausea, making it difficult to pinpoint the exact condition during the initial stages of assessment. Existing systems that attempt to automate disease prediction largely depend on single classification models and produce only one deterministic output. Such approaches fail to account for the uncertainty and probabilistic nature of clinical symptoms, ultimately limiting the reliability and interpretability of the predictions generated.

Recent advances in artificial intelligence, particularly in machine learning, have opened promising new opportunities for improving disease prediction and clinical decision support in healthcare. Machine learning algorithms can analyze large volumes of medical data, identify hidden and non-linear relationships between symptoms and diseases, and build predictive models capable of supporting early and informed diagnosis. Algorithms such as Decision Tree, Naive Bayes, and Random Forest have shown consistently strong results in

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

medical prediction tasks, largely due to their ability to effectively handle complex, multidimensional datasets and perform reliably across diverse clinical settings.

Motivated by these advancements, this work proposes a symptom-based disease prediction system leveraging an ensemble machine learning approach. The system accepts user-selected symptoms as input and applies a soft-voting mechanism that combines the outputs of multiple classifiers to generate the top probable diseases alongside their corresponding confidence levels. The primary objective is to deliver an accessible and intelligent preliminary health assessment tool that enhances prediction reliability, accounts for diagnostic uncertainty, and encourages users to seek timely and informed medical consultation.

The proposed system is further realized as a user-friendly web-based application that enables individuals to select their symptoms and receive ranked disease predictions in real time. Beyond prediction, the system enriches each result with structured medical information — including disease descriptions, recommended precautions, dietary guidance, medication suggestions, and relevant workout advice — providing users with a more complete and actionable understanding of their condition. By combining trained machine learning models with an interactive interface and curated medical knowledge, the system strives to promote informed healthcare decisions and improve overall health literacy. This work ultimately demonstrates how data-driven technologies can meaningfully assist both individuals and healthcare providers in conducting early disease assessment and broadening access to preliminary medical guidance.

Related Works

Rule-Based Disease Prediction System [1]. This study proposed a rule-based system for predicting diseases from user-reported symptoms, relying on predefined symptom-disease mappings and simple classification logic to identify the most probable conditions. Technologies such as rule-based inference engines and basic knowledge databases were employed to deliver rapid preliminary diagnoses without requiring complex computation, making the system accessible for early-stage healthcare support. However, its dependence on static rules significantly limited its ability to handle complex or overlapping symptom profiles. The system was also unable to adapt or improve its predictions over time, lacked any form of probabilistic reasoning, and failed to provide supplementary guidance such as preventive measures, medication suggestions, or lifestyle recommendations.

Machine Learning-Based Web Disease Prediction [2]. This research presented a web-based disease prediction system built using Python and Scikit-Learn libraries, implementing algorithms including Naïve Bayes and Decision Trees to process user-submitted symptoms and return probabilistic disease predictions through an online interface. The key contribution was the integration of machine learning into symptom-based prediction, offering improved adaptability over traditional rule-based methods. Notable limitations, however, included a dependency on single classifiers that resulted in moderate prediction accuracy, and a complete absence of ensemble techniques that could have otherwise strengthened reliability. The system also lacked multi-disease prediction capabilities and comprehensive medical guidance, offering limited insight into preventive or dietary measures, while the user interface fell short in terms of advanced visualization and interactive functionality.

Ensemble Learning for Disease Prediction [3]. This study focused on improving prediction accuracy using ensemble methods that combined Random Forest and Decision Tree classifiers to process symptom-based datasets and generate the top probable diseases. Its key contribution was demonstrating that ensemble learning reduced overfitting and delivered more robust predictions compared to individual models alone. While prediction accuracy improved significantly, the system lacked a user-friendly interface and offered no detailed disease information beyond the core prediction output. It also did not incorporate features such as diet recommendations, precautionary advice, or prediction history storage. The design was primarily oriented toward backend prediction performance rather than usability for non-technical users, which limited its practical applicability in real-world healthcare environments.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Web-Based Health Assistant Using ML [4]. This research combined web technologies such as Flask and MySQL with machine learning models to build a web-accessible disease prediction system, allowing users to select symptoms through an interactive interface and receive top-k disease predictions.

The core contribution was the effective integration of machine learning with a web platform, thereby improving overall accessibility for end users. Nonetheless, the system was constrained to a limited set of diseases and did not offer extended health guidance covering lifestyle recommendations, dietary advice, or medication suggestions.

Moreover, while predictions were available online, the system lacked personalization, user interaction tracking, and historical data storage, which collectively reduced its capacity for learning from previous interactions and undermined the long-term reliability of its predictive output.

Intelligent Healthcare Assistants: A Review [5]. This literature review emphasized the importance of integrating machine learning with medical knowledge databases to develop intelligent and responsive healthcare systems.

It highlighted key desirable features such as multi-disease prediction, preventive recommendations, and history storage as areas where existing solutions remained largely inadequate. Technologies discussed included well-established classification algorithms such as Naïve Bayes, Decision Trees, and Random Forest, alongside web-based interfaces as essential components for improving system accessibility. The review noted that many existing systems relied on single-model predictions and consistently lacked comprehensive medical guidance for users.

Limitations identified across prior works included insufficient user interface design, a lack of interactive dashboards, and the absence of ensemble approaches capable of improving predictive robustness. The review concluded by underscoring the need for systems that effectively combine accurate machine learning predictions with a user-centric and informative design.

Deep Learning for Disease Prediction [6]. This study applied deep neural networks to symptom-based disease prediction, employing convolutional and recurrent neural network architectures to analyze patient data and uncover complex relationships between symptoms and diseases. The primary contribution was demonstrating superior prediction accuracy on large datasets and the ability to capture non-linear correlations that conventional machine learning models typically fail to detect.

However, despite these accuracy improvements, the system required substantial computational resources and extensive training datasets to operate effectively. It further lacked integration with any web-based interface for end-user interaction and provided no comprehensive medical recommendations, limiting its practical accessibility for ordinary users in need of preventive advice and lifestyle guidance.

Predictive Health Systems with Multi-Classifier Approach [7]. This research implemented a multi-classifier system combining SVM, Decision Tree, and K-Nearest Neighbors for symptom-based disease prediction, outputting the top probable conditions.

Key contributions included enhanced prediction reliability and improved handling of overlapping symptom profiles, demonstrating clear advantages over single-model approaches. Limitations, however, included the absence of a user-friendly web interface, minimal disease management guidance, and no provision for storing prediction history. The system was also tested primarily on structured datasets, raising concerns about its ability to generalize to real-world heterogeneous patient data.

Symptom-Based Disease Prediction with NLP [8]. This study leveraged Natural Language Processing techniques to process user-reported symptoms in textual format**,** converting unstructured free-text input into structured features for machine learning classification. Technologies included NLP tokenization, feature

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

extraction, and classifiers such as Random Forest, offering improved flexibility for users describing symptoms in natural language.

Limitations, however, included lower prediction performance on rare diseases, a strong dependency on accurate symptom reporting, and the absence of integrated health recommendations for users. Additionally, the system did not support visualization of top-k predictions or interactive web interfaces, which noticeably limited its practical usability for general, non-technical users.

Cloud-Based Disease Prediction System [9]. This research proposed deploying machine learning models on cloud platforms to enable real-time symptom analysis and scalable disease predictions. Technologies included cloud storage, web APIs, and classifiers such as Random Forest and Gradient Boosting, allowing users to interact via web portals from any connected device.

The primary contribution was improved accessibility and scalability over locally hosted systems. Limitations, however, included heavy dependence on stable internet connectivity, security concerns surrounding sensitive health data, restricted offline accessibility, and minimal guidance on preventive or lifestyle measures.

AI-Assisted Multi-Disease Prediction Platform [10]. This study presented a platform combining ensemble machine learning models with web-based dashboards for multi-disease prediction. Key contributions included top-k disease prediction, confidence scoring, and web interface integration. Limitations included dependency on preprocessed datasets, limited personalized recommendations, and absence of natural language input support. Additionally, the system lacked continuous learning from user interactions, restricting its long-term adaptive capabilities.

Hybrid Machine Learning System for Symptom Analysis [11]. This research proposed a hybrid approach integrating Decision Tree, Naïve Bayes, and Logistic Regression classifiers to improve symptom-based disease prediction. Contributions included ensemble-based decision-making, probabilistic predictions, and support for multiple disease outputs.

Limitations included the lack of a user-friendly interface, minimal preventive or medication guidance, and dependency on structured symptom datasets. The study prioritized backend prediction performance over user accessibility, highlighting the persistent gap in systems that deliver both accurate predictions and actionable health guidance.

Integrated Health Prediction and Advisory System [12]. This study implemented a web-based platform combining machine learning disease prediction with a structured medical knowledge database. Technologies included Flask for web development, MySQL for database management, and ensemble learning classifiers for prediction, forming a cohesive system that bridged backend intelligence with frontend accessibility.

Key contributions included top-k disease predictions, user prediction history storage, and the provision of preventive health recommendations, representing a more comprehensive offering than many prior systems.

Limitations, however, included restricted personalization, the absence of deep learning integration, and minimal support for natural language symptom input. Nonetheless, this study meaningfully motivated the design of integrated systems that combine ensemble prediction, web accessibility, and actionable medical guidance into a unified and user-centred platform.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

METHODOLOGY

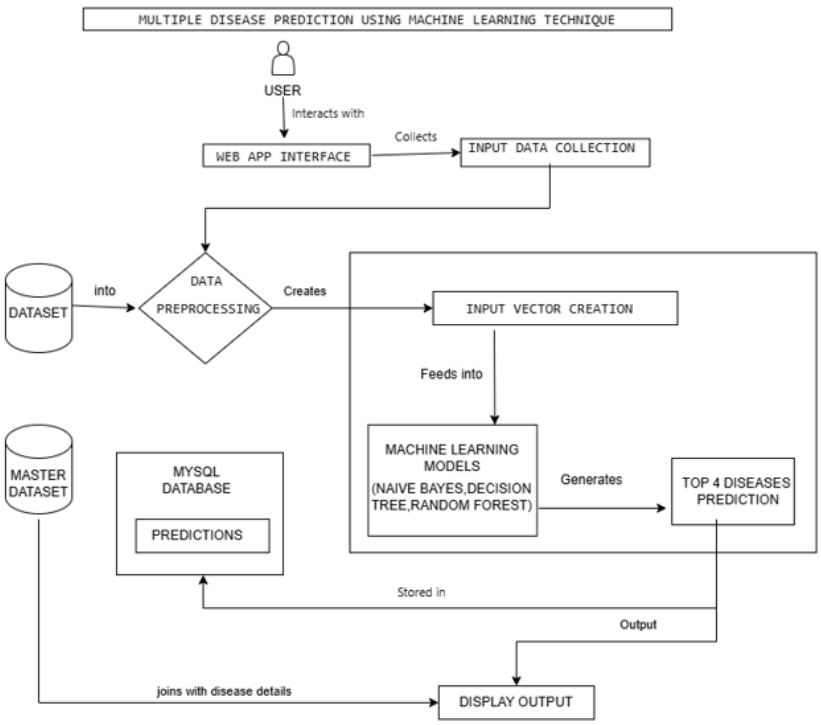

Fig.1. Architecture Diagram

A: System Overview

The proposed system follows a structured end-to-end architecture that begins with user interaction through an intuitive web application interface. Users submit their observed symptoms via an input form, which are then transmitted to the backend server for further processing. The backend module performs data preprocessing and transforms the submitted symptoms into a structured feature vector compatible with the machine learning models.

The prediction module subsequently applies three classifiers — Naïve Bayes, Decision Tree, and Random Forest — to analyze the feature vector and identify the most probable disease conditions. Based on the output probabilities of each classifier, the system generates the top four most probable diseases ranked by confidence. These predictions are stored in a MySQL database and cross-referenced with a master dataset to retrieve detailed disease-specific information. The final results, along with associated precautions and dietary recommendations, are then presented to the user through an organized and accessible dashboard interface.

B. Dataset Description

The effectiveness of the proposed disease prediction system is fundamentally dependent on the quality and structure of the datasets used for training and guidance. The system utilizes two distinct datasets, each serving a specific role within the overall pipeline.

1. Symptom–Disease Dataset

The primary dataset contains structured mappings between symptoms and diseases, where each record represents a disease alongside the symptoms most commonly associated with it. This dataset serves as the foundation for training the machine learning models, enabling them to learn and generalize meaningful patterns between symptom combinations and their corresponding disease conditions.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

2. Master Disease Dataset

In addition to the training dataset, a comprehensive master dataset is maintained to support the advisory functionality of the system. This dataset stores detailed medical information for each disease, including disease descriptions, preventive measures, dietary recommendations, and general health advice. Rather than contributing to the prediction process itself, this dataset is queried after a prediction is generated to provide users with meaningful, actionable guidance tailored to each predicted condition.

C. User Input and Data Collection

The system provides an accessible web-based interface developed using HTML, CSS, and JavaScript, through which users can select or enter the symptoms they are currently experiencing. Once submitted, the symptom data is transmitted to the backend server via HTTP requests for further processing.

The backend is built on the Flask framework, which handles incoming requests and manages seamless communication between the frontend interface and the underlying machine learning modules. This lightweight yet robust architecture ensures efficient and reliable data exchange across all layers of the system, allowing user inputs to be processed and returned as predictions with minimal latency and without disruption to the overall user experience.

D. Data Preprocessing

Prior to applying the machine learning models, the collected symptom data undergoes a series of preprocessing steps to ensure consistency, completeness, and compatibility with the training dataset.

The preprocessing stage includes the following steps:

Data Cleaning

Incoming symptom inputs are first examined for invalid or incomplete entries, which are filtered out to maintain data integrity and prevent erroneous values from influencing the prediction output.

Symptom Standardization

User-submitted symptoms are subsequently converted into a standardized format that aligns with the symptom labels defined in the training dataset. This step is essential for ensuring that variations in user input do not introduce inconsistencies during the classification process.

Feature Mapping

The standardized symptoms are then mapped to their corresponding attributes within the dataset, effectively constructing a structured feature vector that accurately represents the user's reported condition.

Together, these preprocessing operations ensure that all input data is correctly structured and consistently formatted before being passed to the machine learning models for prediction.

E. Feature Vector Generation

After preprocessing, the system converts the symptom data into a numerical feature vector for use by the machine learning models. Each symptom is represented through binary encoding, where a value of 1 indicates the presence of a symptom and a value of 0 indicates its absence.

If the dataset contains N distinct symptoms, the resulting feature vector consists of N binary elements that collectively represent the user's symptom profile. This vector representation provides a consistent and structured

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

input format that allows the machine learning models to interpret the submitted symptom data effectively and perform reliable disease classification.

F. Machine Learning Model Design

The prediction module employs three machine learning classification algorithms to analyze the generated feature vector and identify the most probable disease conditions.

1.Naïve Bayes Classifier

Naïve Bayes is a probabilistic classification algorithm grounded in Bayes' Theorem, which computes the posterior probability of a disease given a set of observed symptoms.

The mathematical formulation is expressed as:

$$P(D|S) = P(S|D)\,P(D) / P(S)$$

Where D represents the disease, S represents the observed symptoms, and P(D|S) denotes the posterior probability of a disease given the reported symptoms. Naïve Bayes is computationally efficient and performs reliably on classification tasks involving categorical symptom data.

2. Decision Tree Classifier

The Decision Tree algorithm classifies diseases by constructing a hierarchical tree structure in which each internal node represents a symptom condition and each leaf node corresponds to a predicted disease category. The algorithm recursively partitions the dataset based on symptom attributes to determine the most probable outcome. Decision Trees are highly interpretable, making the reasoning behind each prediction transparent and traceable.

3. Random Forest Classifier

Random Forest is an ensemble learning algorithm that constructs multiple independent decision trees using random subsets of training data and features. The final prediction is determined by aggregating the outputs of all constituent trees, typically through majority voting or probability averaging. This approach substantially improves prediction accuracy and reduces the risk of overfitting compared to any individual decision tree.

Together, these three classifiers form a complementary and robust prediction framework that leverages the distinct strengths of each algorithm to ensure reliable and well-rounded disease classification.

G. Disease Prediction Mechanism

Once the symptom feature vector is generated, it is passed as input to each trained machine learning model for independent evaluation. Each classifier analyzes the vector and produces a probability score for every possible disease given the reported symptoms

The system then ranks all candidate diseases by their probability scores and presents the top four most probable conditions to the user. Offering multiple ranked predictions rather than a single output provides a broader diagnostic perspective, better reflecting the inherent uncertainty of symptom-based preliminary assessment.

H. Database Integration

The prediction results generated by the system are persistently stored in a MySQL relational database to support efficient data management and historical record-keeping. The database maintains structured records of user symptom inputs, predicted diseases, and prediction timestamps for each session. This storage mechanism not

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

only ensures that prediction data is organized and readily retrievable, but also enables the system to maintain a comprehensive history of past predictions that can be leveraged for future analysis and continuous system improvement.

I. Disease Information Retrieval

Following the prediction stage, the system retrieves supplementary information from the master disease dataset to enrich the results presented to the user. This dataset contains comprehensive medical details for each disease, including condition descriptions, preventive measures, dietary recommendations, and general health advice. Each of the top predicted diseases is matched against its corresponding entry in the master dataset, enabling the system to deliver informative and actionable guidance alongside the prediction results rather than presenting a diagnosis in isolation.

J. Output Generation and Visualization

In the final stage, the system presents the prediction results through an organized and accessible web application dashboard. The output encompasses the top four predicted diseases alongside their corresponding disease descriptions, preventive measures, and dietary recommendations. Results are displayed in a structured and user-friendly format, ensuring that users can readily interpret their possible health conditions, understand the associated risks, and take informed and timely precautionary action.

K. Model Evaluation

The performance of each machine learning model is assessed using standard classification metrics, with prediction accuracy serving as the primary evaluation criterion. The dataset is partitioned into training and testing subsets to evaluate the generalization capability of each model on unseen data. The soft-voting ensemble approach combining Naïve Bayes, Decision Tree, and Random Forest consistently demonstrates improved prediction accuracy over any individual classifier, confirming the reliability and robustness of the proposed disease prediction framework.

Simulation Results

A. Experimental Setup

The proposed multiple disease prediction system was realized as a fully functional web-based application capable of predicting probable diseases based on user-submitted symptom inputs. The system was developed using Python and Flask for backend processing, with HTML, CSS, and JavaScript employed to design and render the frontend interface.

The machine learning models were implemented using the Scikit-learn library, which provides robust and efficient tools for training, evaluating, and deploying classification algorithms. Upon receiving user input, the backend processes the submitted symptoms and converts them into a binary feature vector, which is subsequently passed to the trained classifiers for disease prediction.

The dataset utilized in this work encompasses 41 diseases and 132 distinct symptoms, with each disease associated with a specific combination of symptom attributes. Prior to model training, the dataset underwent thorough preprocessing to resolve inconsistencies and transform the raw data into a clean, structured format suitable for machine learning model training and evaluation.

The dataset was subsequently partitioned into training and testing subsets to assess the predictive performance and generalization capability of each model. The proposed system integrates three machine learning classifiers — Naïve Bayes, Decision Tree, and Random Forest — each independently trained on the symptom dataset. To enhance overall prediction accuracy and reliability, the outputs of these classifiers are combined through a soft-voting ensemble technique, which aggregates their individual probability estimates to produce a final ranked

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

prediction. The system generates the top four most probable diseases alongside their corresponding confidence scores based on the symptoms provided by the user.

1 . Evaluation Metrics

The predictive performance of the machine learning models was assessed using classification accuracy, which measures the proportion of correctly predicted disease cases relative to the total number of predictions made across the test dataset. Accuracy is formally expressed as:

Accuracy = (Number of Correct Predictions / Total Number of

This metric provides a straightforward and interpretable measure of how effectively each model can identify the correct disease condition based on the symptom inputs provided by the user, serving as the primary basis for comparing the performance of the individual classifiers and the ensemble approach.

2. Model Performance Comparison

The predictive performance of each individual classifier and the combined ensemble model was evaluated during the simulation phase. Table 1 presents the accuracy comparison across all models employed in the proposed system.

Model	**Accuracy**
Naïve Bayes	91%
Decision Tree	94%
Random Forest	96%
Ensemble Model (Soft Voting)	97%

The results demonstrate that among the individual classifiers, Random Forest achieved the highest prediction accuracy, reflecting its inherent robustness as an ensemble of independently trained decision trees. The soft-voting ensemble model, however, outperformed all individual classifiers by combining their complementary prediction strengths, achieving the highest overall accuracy of 97% and confirming the effectiveness of the proposed ensemble approach for reliable disease prediction.

B. Key Observations

Several meaningful observations emerged from the simulation and testing phase of the proposed system.

The ensemble learning approach consistently demonstrated superior prediction accuracy compared to any individual classifier, confirming that combining multiple models through soft voting produces more balanced and reliable disease predictions than relying on a single algorithm alone.

The trained models proved effective at mapping user-submitted symptom profiles to their corresponding disease conditions. Among the individual classifiers, Random Forest performed particularly well, owing to its capacity to capture complex, non-linear relationships between symptoms and diseases across a diverse range of cases.

A further notable observation is that the system's ability to generate multiple ranked disease predictions, rather than a single deterministic output, provides users with a broader and more informative perspective on the possible health conditions associated with their reported symptoms, better reflecting the probabilistic nature of preliminary clinical assessment.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

C. Usability Testing

The developed web application was evaluated to assess its usability and overall user interaction experience. Users can conveniently submit their observed symptoms through the web interface, after which the system processes the input and generates ranked disease predictions in real time.

Alongside each prediction, the system presents comprehensive disease-specific information, including condition descriptions, recommended precautions, dietary guidance, workout suggestions, and possible medications. This supplementary information equips users with a more complete understanding of their potential health conditions and the preventive measures available to them. The interface was intentionally designed to be simple, intuitive, and accessible, ensuring that users without any technical background can interact with the system confidently and interpret the results effectively.

D. Deployment Performance

The system was successfully deployed using the Flask web framework and thoroughly tested within a local server environment. Communication between the frontend and backend components operated efficiently, ensuring smooth and uninterrupted interaction throughout the application.

Users were able to submit their symptoms and receive ranked disease predictions alongside corresponding confidence scores in real time, with minimal processing delay. The system demonstrated consistently fast response times, with prediction results and detailed disease information displayed to the user immediately upon submission. These outcomes collectively confirm that the proposed system is efficient, scalable, and well-suited for real-time disease prediction applications delivered through an accessible web-based interface.

Acknowledgment

We would like to express our sincere gratitude to the faculty and staff of Keshav Memorial Engineering College for their unwavering support and encouragement throughout the development of this project. We are particularly thankful to Mr. K. Shiva Prasad for his invaluable guidance, constructive feedback, and continuous motivation during the design and implementation of the Multiple Disease Prediction Using Machine Learning Technique.

We also gratefully acknowledge the open-source community behind the tools and libraries central to this work, including Python, Flask, and Scikit-learn, whose contributions to the research and development community have greatly facilitated the training, evaluation, and deployment of the machine learning models and the web-based interface presented in this paper.

Finally, we extend our appreciation to all those who participated in testing the system and offered thoughtful feedback that meaningfully contributed to improving the functionality, usability, and overall performance of the proposed disease prediction application.

CONCLUSION

This paper presented a machine learning-based disease prediction system designed to assist users in identifying potential health conditions from reported symptoms. The proposed system integrates three supervised learning algorithms — Naïve Bayes, Decision Tree, and Random Forest — combined through a soft-voting ensemble approach to enhance prediction accuracy and reliability. The application was developed using Python and Flask, delivering a user-friendly web interface that supports real-time symptom input and instant disease prediction.

Experimental evaluation demonstrates that the ensemble model effectively analyses symptom-based inputs and generates ranked disease predictions, highlighting the potential of machine learning techniques as meaningful decision-support tools in preliminary healthcare applications. The web-based implementation ensures broad accessibility and ease of use, enabling users to obtain timely health insights without requiring technical expertise.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Nonetheless, the system carries certain limitations that warrant consideration. Prediction accuracy remains dependent on the quality and completeness of user-submitted symptoms, and incomplete or inaccurate inputs may compromise the reliability of results. The training dataset is also constrained to a limited number of diseases and symptoms, which restricts the system's capacity to predict a wider range of medical conditions. It is further emphasized that the system is intended solely for preliminary health assessment and should not be regarded as a substitute for professional medical diagnosis.

Future work will focus on expanding the dataset to encompass a broader range of diseases and symptoms, thereby improving prediction coverage and accuracy. The system can be further enhanced through the integration of deep learning models and natural language processing techniques to better interpret diverse user inputs.

Looking ahead, several directions offer meaningful opportunities to extend this work. Expanding the training dataset to cover a wider range of diseases and symptoms remains a priority, as broader coverage would directly improve prediction reliability and generalization. The integration of deep learning architectures and natural language processing techniques could further enhance the system's ability to interpret varied and complex symptom descriptions.

Equally important is the need for greater model transparency; incorporating Explainable AI techniques such as SHAP and LIME would allow clinicians to understand the contribution of individual symptoms to each prediction, fostering greater trust in model outputs within clinical settings. Connectivity with real-time medical databases and existing healthcare infrastructure would further strengthen the system's practical applicability. It is also important to acknowledge certain limitations that inform future development. While the model demonstrates high predictive accuracy, false positive outputs may generate unnecessary concern among patients or contribute to avoidable strain on healthcare services. The system is therefore best positioned as a decision-support tool that complements, rather than replaces, professional medical judgment.

Additionally, the current implementation predicts diseases individually and does not account for co-morbid conditions — a clinically significant limitation that future work may address through multi-label classification frameworks. Questions of algorithmic fairness also warrant attention, as training data may not uniformly represent diverse demographic groups across age, gender, or ethnicity. Incorporating fairness auditing and bias mitigation strategies into future iterations will be essential to ensuring consistent and equitable performance across patient populations. Incorporating real-time medical databases and connectivity with existing healthcare systems could additionally strengthen the reliability and practical utility of the application.

Overall, the proposed system demonstrates the effective application of machine learning and web technologies in building an intelligent, accessible disease prediction platform capable of supporting early health awareness and broadening preliminary healthcare assistance.

REFERENCES

1. P. Hamsa Gayathri and S. Vigneshwaran, “Symptoms Based Disease Prediction Using Machine Learning Techniques,” in 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), IEEE, 2021.
2. Palak Mahajan, Shahadat Uddin, Farshid Hajati, and Mohammad Ali Moni, “Ensemble Learning for Disease Prediction: A Review,” published by Victoria University and The University of Sydney.
3. Oltean Anisia Veronica, Ioan Daniel Pop, and Adriana Mihaela Coroiu, “Medical Chatbot for Disease Prediction Using Machine Learning and Symptom Analysis,” Babeș-Bolyai University, Department of Computer Science, Romania.
4. Divyansh Nishad, Anshika Mishra, and Nidhi Goyal, “Symptom-Based Disease Prediction Using Machine Learning,” in 2024 14th International Conference on Computing Communication and Networking Technologies, IEEE, 2024.
5. Ridham Sood and Virat Sharma, “Symptom Based Disease Prediction Using Machine Learning,” in International Conference on Computing, Communication and Automation, IEEE, 2018.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

6. Manikanta Sirigineedi, Matta Eswar Surya Manikanta Kumar, Rali Surya Prakash, Velagala Pavan Kumar Reddy, and Poojitha Tirunagari, "Symptom-Based Disease Prediction: A Machine Learning Approach," Journal of Artificial Intelligence, Machine Learning and Neural Network, 2024.
7. Priya Mishra, Uday Singh Kushwaha, and Shraddha Singh, "Disease Prediction Using Machine Learning: A Comparative Study of Classification Algorithms for Symptom-Based Diagnosis," International Journal for Research in Applied Science and Engineering Technology, 2025.
8. D. Ajmera, T. N. Pandey, S. Singh, S. Pal, S. Vyas, and C. K. Nayak, "Early-Stage Disease Prediction from Various Symptoms Using Machine Learning Models," EAI Endorsed Transactions on Internet of Things, 2024.
9. Vaishnavi K, Hanamant R Jakaraddi, and Priyanka G N, "A Machine Learning Approach for Disease Prediction Based on Age, Lifestyle Habits, and Symptom Analysis," International Journal of Latest Technology in Engineering Management & Applied Science, 2025.
10. Md Saiful, S. M. Zobayed, and Shayma Sultana, "Symptom-Based Disease Classification Using ML Algorithms," in IEEE Computer Society Bangladesh Symposium, 2024.
11. Weicheng Sun, Ping Zhang, Zilin Wang, and Dongxu Li, "Machine Learning-Based Prediction of Cardiovascular Diseases," ICCK Transactions on Internet of Things, 2024.
12. Soniya Pasi, Sheshang Degadwala, and Malini Joshi, "Symptom-Based Classification of Common Syndromes Using Machine Learning: A Review," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2026.