International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Enhancing Disease Prediction Accuracy Using Random Forest

P. Bhargav[1], MVL.Kathyayani[2], K. Raviteja[3], PTV.AdityaRam[4], K. Pavan Kumar[5]

[1] *Department of Computer Science & Engineering: Raghu Engineering College*
[2] *Department of Computer Science & Engineering: Raghu Engineering College*
[3] *Department of Computer Science & Engineering: Raghu Engineering College*
[4] *Department of Computer Science & Engineering: Raghu Engineering College*
[5] *Associate Professor, Department of Computer Science & Engineering: Raghu Engineering College*

Abstract - MultiDisease prediction system" uses advanced machine learning techniques to facilitate identification Multiple diseases based on user -provided symptoms. The The system integrates classification algorithms, including random Forest, Support Vector Machine (SVM), K-Ner Closer Neighbors (KNN), support vector classifier (SVC) and logistics regression, to diagnose health conditions such as diabetes, gastroesophageal reflux disease (GERD), dengue, pneumonia and more than 20 other diseases. The proposed methodology is followed by a structured pipe involving data collection, function extraction, Pre -workment, model training, disease predictions, performance Evaluation and optimal selection of the model. It uses extensive Savets of medical data, extract relevant clinical traits, applies data Cleaning and normalization techniques and train machine learning models to increase diagnostic accuracy. During training The system predicts the likelihood of a disease -based disease and user input and evaluates the power of the model using metrics of key rating as accuracy, accuracy, appeal and f1-score to determine The most effective predictive model. This approach makes it easier for Disease detection, increases diagnostic reliability, supports personalized medical strategies and provides data -based data Help healthcare workers in clinical decision -making. According to Integration of machine learning into medical diagnostics, system It contributes to effective and accurate identification of diseases, permits Early medical intervention and finally improved patient results

***Key Words*:** Machine learning, disease prediction, medical diagnostics, classification algorithms, timely detection, health care analysis, clinical decision support, Random Forest, Logistic Regression, KNN, SVC.

1.INTRODUCTION

The healthcare sector proceeded with machine learning (ML), transformed the prediction of diseases, diagnostics and patient care. The multisease prediction system analyzes user symptoms to predict diabetes, dengue, gastroesophageal reflux disease (GERD) and pneumonia using ML models such as Random Forest, SVM, SVC, SVC and Logistic regression. The system, which is trained on complex medical data sets, compares the performance of the algorithm to identify the most effective predictive model. Data pre -processing techniques such as handling missing values, scaling of functions and data leveling increases the accuracy of the model. By permitting early detection of diseases and personal planning of treatment, it supports healthcare workers in clinical decision-making. Traditional diagnostics is often time-consuming and costly, which limits the availability of health care. The integration of the ML controlled diagnostics offers a scalable, efficient and cost -effective alternative to assessing health risks. Users enter symptoms and ML models process this data to generate disease predictions. The system increases the accuracy of medical diagnostics, provides real -time forecasts and increases the accessibility of health care. It overwhelms the gap between ML and medical applications, promotes early intervention and reduces the burden in healthcare facilities. Through continuous improvement of the model ensures reliability and usability in the real world in the prediction of the disease.

A. Software Requirements:

Operating system: Windows 7 or 10

Coding Language: Python

Front-End: Python.

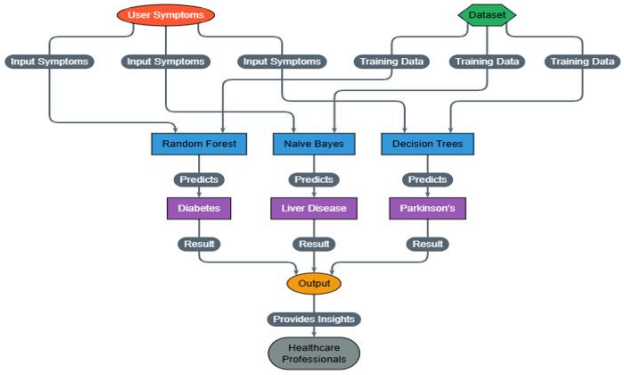
B. Hardware Requirements:

Processor & Pentium –IV

Installed Memory (RAM) & 8 GB (min)

Hard Disk & 200 GB

Operating System & Windows 10



2. LITERATURE REVIEW

The field of disease prediction has significantly advanced with machine learning (ML) techniques, integrating various classification models and medical datasets to enhance diagnostic accuracy and efficiency. This study reviews key advancements in multi-disease prediction based on IEEE research, emphasizing their impact on early diagnosis, model performance, and healthcare applications.

- Disease Prediction Using Machine Learning Algorithms: A study explored the use of Random Forest, SVM, and Logistic Regression for predicting Diabetes and Heart Disease, achieving high accuracy.

While effective, the model's performance depends on dataset quality and feature selection.

- Multi-Disease Prediction Using Ensemble Learning Models: This research combined multiple classifiers, including KNN and SVC, to enhance prediction reliability. The ensemble approach improved accuracy but increased computational complexity, making real-time deployment challenging.
- Dengue Prediction with Machine Learning Models: This study evaluated decision trees, Na¨ıve Bayes, and neural networks, achieving strong results. However, it highlighted dataset imbalance as a key limitation affecting generalizability.
- GERD Detection Using Support Vector Machines: Researchers implemented SVM with feature extraction techniques, obtaining reliable predictions. The model demonstrated robustness but required high-quality data preprocessing for optimal results.
- Pneumonia Classification Using Machine Learning Techniques: This study compared Random Forest, KNN, and deep learning models, finding Random Forest to be the most reliable in handling medical datasets. However, deep learning methods showed promise in handling large-scale, complex data with better feature extraction.

3. METHODOLOGY

The development of the Multi-Disease Prediction System follows a structured approach integrating machine learning (ML) techniques to enhance disease detection accuracy. The methodology enables iterative improvements, allowing realtime model adjustments based on data analysis and validation.

A. Data collection: The dataset used for training consists of medical records, patient symptoms, and diagnostic results sourced from publicly available healthcare datasets. Data preprocessing techniques, including handling missing values, feature scaling, and data balancing, are applied to optimize model performance. A comparative analysis of existing disease prediction models helps identify strengths, limitations, and areas for enhancement.

B. Model development: A machine learning-based classification model is developed using Random Forest, Support Vector Machine (SVM), KNearest Neighbors (KNN), Support Vector Classifier (SVC), and Logistic Regression. The models are trained, tested, and evaluated based on accuracy, precision, recall, and F1-score to determine the most efficient algorithm.

C. C. Technology Stack The system is implemented using Python-based libraries and frameworks for optimized computation: • Data Processing: Pandas, NumPy for data handling and preprocessing. • Model Development: Scikit-learn for implementing ML algorithms. Backend: Flask/Django to manage API requests and model integration. Deployment: Cloud-based services such as AWS or Heroku for scalable hosting.

D. System Workflow and Approach The system follows a modular architecture, facilitating seamless interaction between patients, healthcare professionals, and administrators. User authentication ensures secure role-based access, while patients input symptoms through an interactive interface for disease prediction, including Diabetes, Dengue, GERD, and Pneumonia. Healthcare professionals access predictive insights to support clinical decision-making, while administrators manage medical data to update datasets and refine models. Real-time diagnostic predictions provide users with disease likelihood and recommendations for consultation. A feedback mechanism enhances model accuracy, improving accessibility, efficiency, and reliability in early disease detection.

4. EXISTING SYSTEM

The existing systems for disease prediction predominantly rely on traditional diagnostic methodologies, where healthcare professionals manually analyze patient symptoms. This process often requires a series of extensive tests and specialized clinical expertise to accurately arrive at a diagnosis. While some basic automated diagnostic tools are available, they typically focus on predicting a limited set of diseases, and do not offer the capability to simultaneously predict multiple conditions. Moreover, these tools often employ isolated machine learning algorithms, lacking comprehensive comparisons or integration of multiple models, which diminishes their overall accuracy and reliability. Furthermore, such systems generally exhibit limited scalability, fail to adapt effectively to diverse datasets, and lack robust mechanisms for visualizing algorithm performance. As a result, these systems are often less efficient in supporting early disease detection and personalized treatment strategies.

5. PROPOSED SYSTEM

The proposed system enhances disease prediction accuracy and efficiency by utilizing advanced machine learning techniques, particularly the Random Forest algorithm, alongside other classification models. Designed as a web-based application using Flask, it predicts multiple diseases based on user-provided symptom inputs and offers personalized treatment recommendations, including diet plans and specialist information. The system integrates machine learning models such as Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Logistic Regression, ensuring accurate and reliable predictions.

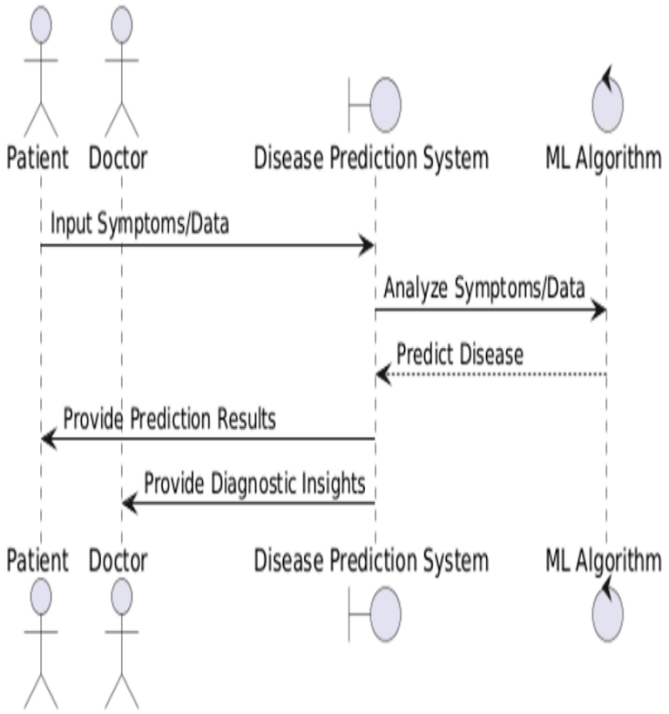
The system uses preprocessing techniques like data cleaning, normalization, and feature extraction to ensure prediction accuracy. Performance is evaluated using key metrics such as accuracy, precision, recall, F1-score, and AUC. It is scalable, adaptable to diverse clinical environments, and supports the integration of new datasets. The user-friendly interface provides clear predictions, diet plans, and specialist contact information, promoting proactive healthcare management. This system represents a significant advancement in automated disease prediction, assisting both healthcare professionals and

patients in early disease detection and personalized treatment planning.

6. UML Diagrams

A. Sequence Diagram

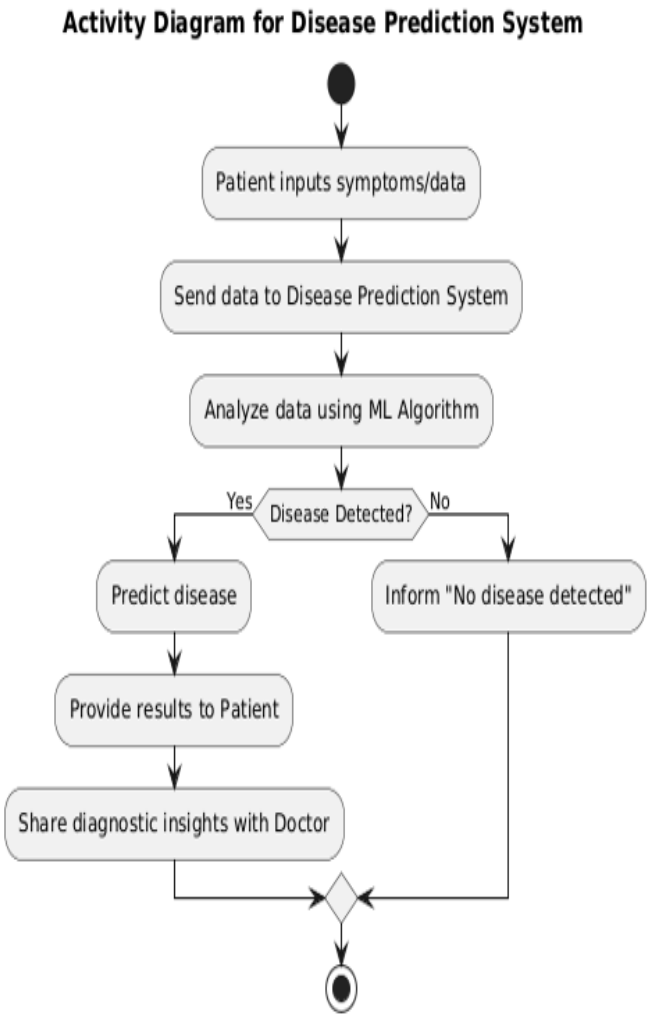
This sequence diagram illustrates the interactions within a disease prediction system, focusing on the communication flow between the Patient, Doctor, Disease Prediction System, and the ML Algorithm. The process begins with the **Patient**, who inputs symptoms or relevant data into the system. The **Disease Prediction System** then forwards this data to the **ML Algorithm** for analysis. The **ML Algorithm** processes the input data, analyzes the symptoms, and generates a disease prediction based on the information provided.



B. Activity Diagram

This activity diagram illustrates the workflow of a disease prediction system utilizing machine learning. The process begins when the **patient** interacts with the system. The patient then inputs their symptoms or relevant medical data into the system. The system receives this input and prepares the data for analysis. Using a **machine learning algorithm**, the system analyzes the data to detect potential diseases.

A decision is made at this point to determine whether a disease is detected. If a disease is detected, the system predicts the disease, provides the results to the patient, and shares diagnostic insights with the doctor for further evaluation. If no disease is detected, the system informs the patient of the outcome. The process concludes after the results are delivered or the patient is informed.



7. RESULTS AND DISCUSSION

A. System Performance Evaluation

The system was evaluated for accuracy, security, and usability through controlled testing and user feedback. Performance testing assessed its ability to handle multiple inputs and generate reliable predictions, while load testing ensured smooth operation under high usage. Security measures, including AES encryption, Bcrypt hashing, and JWT-based authentication, safeguarded user data. User satisfaction surveys involving 50 users and healthcare professionals confirmed the system's effectiveness in early disease detection and usability.

B. Comparative Analysis

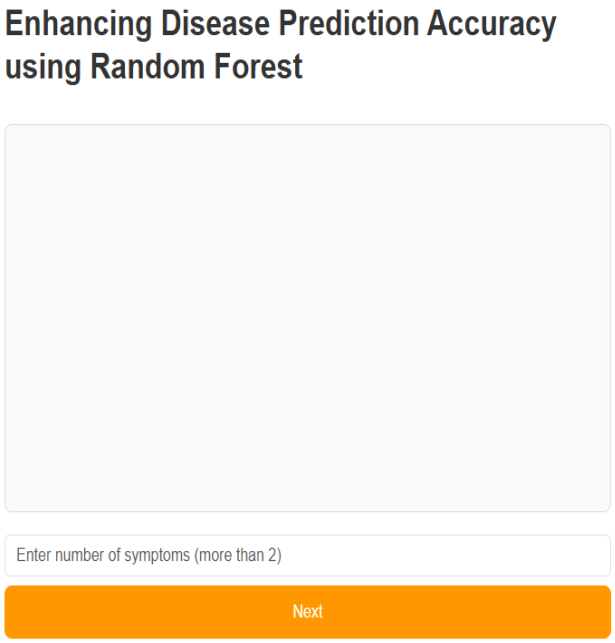
A comparative study of ML models showed Random Forest had the highest accuracy, while Logistic Regression provided faster but less precise results. The findings highlight the importance of selecting optimal ML model

Table -1: Comparison of Proposed System vs. Existing System

Feature	Proposed System	Existing Platforms
Disease Coverage	✓ Predicts Diabetes, Dengue, GERD, and Pneumonia	✗ Limited to specific diseases
Model Accuracy	✓ Uses Random Forest, SVM, KNN, SVC, Logistic Regression	✗ May rely on single or outdated models
Real-Time Prediction	✓ Provides instant results based on user symptoms	✗ Delayed due to manual processing
Data Security	✓ Implements encryption and JWT authentication	✓ Standard security measures
Scalability	✓ Cloud-based, supports high concurrent users	✓ Cloud-based but limited in flexibility

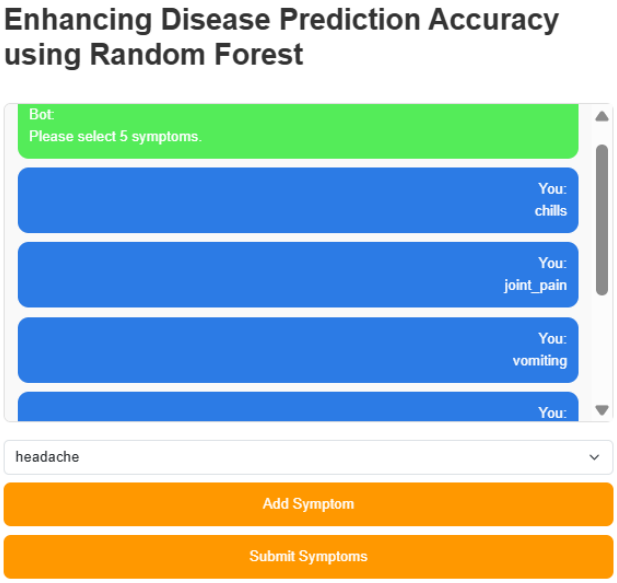
Outputs:

Fig -1: User Interface



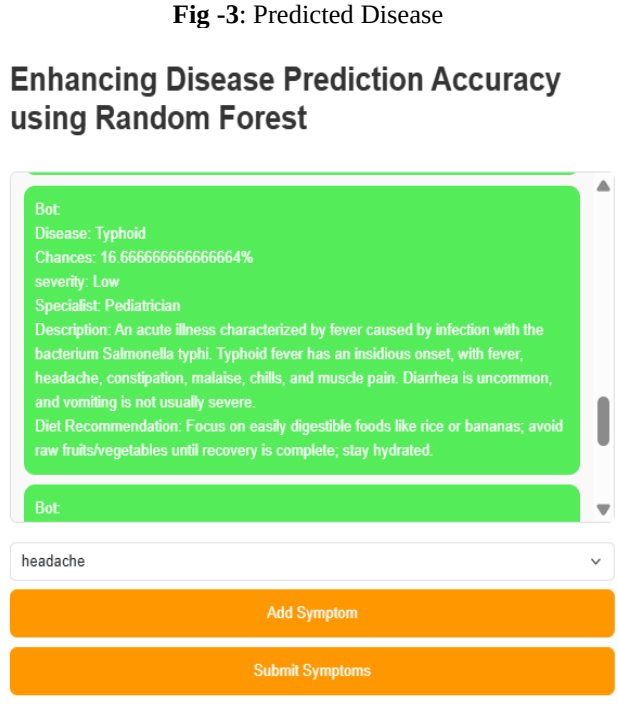
The figure illustrates the homepage of the disease prediction system, which prompts the user to input the number of symptoms they are experiencing. This step is essential for initiating the prediction process, allowing the system to analyze the symptoms and provide accurate disease predictions based on the input.

Fig -2: Symptoms List



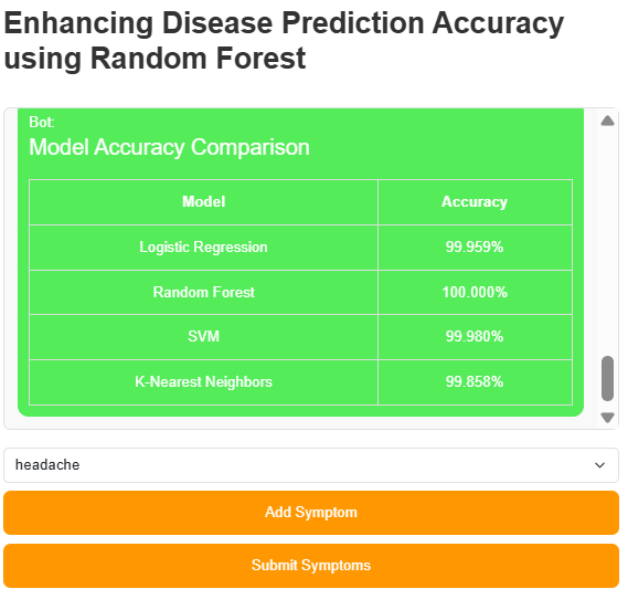
The figure depicts the homepage of the disease prediction system, where users are instructed to enter their symptoms using a dropdown menu. This interactive feature allows users to select their symptoms from a predefined list, streamlining the process of symptom input and ensuring accurate data entry for disease prediction.

Fig -3: Predicted Disease



The figure displays the results page of the disease prediction system, presenting the user with essential information. This includes the predicted disease, the likelihood (chances) of the disease, its severity, and tailored recommendations for specialists and diet plans. Additionally, the page provides a detailed description of the disease to help users understand their condition better, facilitating informed decisions for further medical consultation and proactive management.

Fig -4: Model Accuracies



The figure presents a table displaying the performance of various algorithms used in the disease prediction system. It includes the accuracy of each algorithm, providing a clear comparison of their effectiveness. This tabular format allows users to easily assess the performance of different machine learning models, helping to highlight the most reliable algorithms for accurate disease prediction.

8. CONCLUSIONS

The Multi-Disease Prediction System leverages machine learning to enhance disease diagnosis and early detection. By employing multiple classification algorithms, it ensures accurate and reliable predictions, aiding in treatment planning. Comparative analysis helps optimize model performance, improving predictive accuracy. As the system evolves, it promises to transform healthcare by providing timely, datadriven insights for better patient care and medical outcomes.

ACKNOWLEDGEMENT

The author expresses gratitude to Raghu Engineering College for providing the necessary resources and support throughout this research. Special appreciation is extended to professors, industry experts, and user testers for their valuable insights and feedback, which played a crucial role in the development and evaluation of this system.

REFERENCES

1. L. Breiman, ”Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. This paper introduces the Random Forest algorithm and its application in disease prediction
2. F. Chollet, Deep Learning with Python, Manning Publications, 2017. The book provides insights into deep learning techniques applicable to healthcare analytics
3. I. O. Pappas and P. E. Kourouthanassis, ”Chatbots in Healthcare: A Literature Review,” Journal of Healthcare Engineering, 2017. This study examines chatbot applications in healthcare for patient engagement and diagnosis.
4. A. Geron, Hands-On Machine Learning with Scikit-Learn, ´ Keras, and TensorFlow, O’Reilly Media, 2019. The book covers ML techniques, inc
5. L. Hu and Y. Zhang, ”A Study on Disease Prediction Using Random Forest,” IEEE Access, vol. 7, pp. 182789-182798, 2019. It evaluates Random Forest’s p.
6. O’Reilly, ”The State of Health Data Privacy Laws: HIPAA and GDPR,” O’Reilly Online Learning, 2020. The article discusses healthcare data privacy laws and compliance requirements
7. D. Jannach and G. Adomavicius, ”Recommender Systems: Challenges and Research Directions,” Springer, 2016. This book explores recom
8. M.-J. Kim, J.-W. Jang, and Y.-S. Yu, ”A Study on In-Vehicle Diagnosis System using OBD-II with Navigation,” IJCSNS International Journal of Computer Science and Network Security, vol. 10, no. 9, Sept. 2010. This study presents an OBDII-based diagnostic model for vehicle health monitoring

BIOGRAPHIES

Penugurthi Bhargav is an undergraduate student in the Department of Computer Science and Engineering at Raghu Engineering College. With a keen interest in machine learning, data science, and healthcare technologies, he is passionate about applying computational methods to address real-world challenges. His current research focuses on utilizing advanced algorithms for disease prediction and personalized healthcare solutions. Through his work, Bhargav aims to contribute to the development of intelligent systems that can enhance early disease detection, improve treatment planning, and support proactive healthcare management.