IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, November 2025
Impact Factor: 7.67

AI-Based Disease Prediction And Remedy Recommendation System

Ashwini B G[1] and Shankar B S[2]
Student, Department of MCA[1]
Assistant Professor, Department of MCA[2]
Vidya Vikas Institute of Engineering and Technology, Mysore

Abstract: *A machine learning system is developed to predict diseases using 31 symptom indicators. The project covers data preparation, training a Random Forest model, evaluating its performance, and creating a Flask- based web interface for users to get predictions with a confidence score. It focuses on data quality, model interpretability, deployment, and ethical aspects like bias, privacy, and security. The system provides a complete end- to-end solution, including a clean data pipeline, reproducible training workflow, and evaluation using multiple metrics. It also highlights challenges like limited and imbalanced datasets and suggests improvements such as using larger datasets, better explain ability methods, and secure data handling. Overall, this work demonstrates the practical use of machine learning in identifying diseases at an early stage and serves as a meaningful step toward expanding research and applying such systems in real healthcare settings.*

Keywords: Symptoms, Random Forest, Decision Tree, Machine Learning, Classification

I. INTRODUCTION

Early and accurate disease prediction is a key component of effective healthcare. Symptom-based triage assurances that patients will acquire medical attention, avoid unnecessary delays, and get guidance even when specialists are not immediately available. By converting patient-reported symptoms into structured clinical information, supervised Learning models are able to recognize trends and make recommendations for potential circumstances. This system provides a reliable and efficient tool for offering preliminary medical guidance, supporting early awareness and decision-making for patients.

In this study, a Random Forest model was trained using thirty-one binary symptom attributes collected from patients. The Random Forest algorithm was selected because it manages correlated features efficiently, identifies the most influential symptoms, and remains relatively simple to interpret effectively, identifies the most significant symptoms, and remains relatively easy to interpret. The primary goal is not to replace healthcare professionals but to assist in early triage, enhance patient awareness, and provide meaningful insights that can support clinical decisions effectively complement clinical decision-making, overlapping calls, and variations in vocalization patterns. Humans also tend to find visual identification easier than auditory recognition, although bird calls remain crucial in field surveys where direct sightings are limited.

Considering current healthcare challenges, where doctors often face heavy workloads and limited resources, such intelligent systems offer meaningful support. They deliver quick, cost-effective, and accessible insights, enabling clinicians to make better decisions and guide patients toward timely medical care.

This project follows a structured workflow—from data validation and pre- processing to evaluate, deploy, and train models via a Flask-based web application—demonstrating both the technical feasibility and practical value of AI-driven disease prediction systems.

II. LITERATURE REVIEW

Recently, several of researchers have explored AI-based disease prediction using patient symptom and medical datasets. [1] Presented the Random Forest algorithm, which combines several decision trees to enhance prediction accuracy and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

reduce model variance with feature-level randomness. It reduces over fitting, handles noisy and irrelevant characteristics well, and gives feature importance ratings for structured health data. [2] Discussed machine learning in medical diagnosis, emphasizing pre-processing, handling absent values, and model interpretability to ensure reliable and reproducible results in healthcare applications.

[3] Applied classical ML methods like SVM, k-NN, and ensemble approaches to predict diabetes, showing ensemble methods improve accuracy on imbalanced datasets and highlighting the importance of feature normalization.

[4] Investigated boosting and bagging ensemble methods for symptom-based disease prediction, demonstrating that ensembles handles parse and non linear symptom interactions better than single models, improving classification accuracy.

[5] Focused on liver disease utilizing hybrid feature selection to prediction and pre-processing, showing that selecting relevant features and careful data handling improves accuracy and reduces computation time. [6] Applied Random Forest to breast cancer prediction with small structured datasets, achieving high accuracy and emphasizing hyper parameter tuning and feature scaling for better performance.

[7] Surveyed Explainable AI (XAI) techniques like SHAP and LIME in healthcare, highlighting the importance of interpretability to gain clinical trust. [8] Used deep learning on Electronic Health Records to predict outcomes like mortality, addressing pre-processing, scalability, and interpretability, while showing that advanced models require strong infrastructure.[9]Developed Doctor AI, an RNN- Based method of future prediction diagnoses using sequential HER data, demonstrating that temporal modeling improves prediction of chronic or evolving conditions.

Research indicates that Random and other ensemble methods are effective for disease prediction, as they reduce over fitting, handle noisy features, and identify important symptoms. Classical ML models like SVM work moderately well but struggle with imbalanced datasets, learning, and deep models like RNNs and CNNs can capture complex patterns but need large datasets. Proper data pre-processing, feature selection, and interpretability are essential, with tools like SHAP and LIME improving clinician trust.

Overall, AI- based systems can support early detection, aid doctors, and improve patient outcomes, though challenges like data quality and ethics must be addressed.

III. METHODOLOGY

This project's methodology guarantees that the dataset is properly cleaned, processed, and modeled to produce reliable disease prediction results. Each stage of the process from data validation to model deployment was carefully designed to keep data integrity, fairness, and interpretability, which are essential aspects in healthcare- related machine learning applications.

Data Cleaning:

To maintain dataset quality, a regex-driven parser was implemented for validating the input rows. Each record in the dataset was checked to guarantee that it contained exactly thirty-one binary symptom values (0 or 1) followed by a single categorical disease label. Any rows failing to meet this condition were automatically excluded, thereby preventing potential errors due to missing, extra, or malformed entries. This pre-processing step was essential for preserving dataset consistency and ensuring that only valid, structured data was utilized for model training.

Train/Test Strategy:

Dataset was divided into separate subsets for training and testing to evaluate the model's generalization capability. For disease classes with at least two or more samples, a stratified 80/20 split was applied, maintaining proportional representation of each disease in both subsets. This approach preserved balance and fairness across the data. In cases where certain disease classes contained very few samples, a non-stratified split was used carefully to prevent data leakage and to ensure that test samples remained accessible for evaluation.

Model Choice:

The Random Forest algorithm was selected as the predictive model for this project. Random Forest is particularly effective for healthcare datasets; a site provides a good balance between variance and bias using ensemble learning that combines random feature selection with bagging. The algorithm can efficiently handle binary features, capture non-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

linear relationships among symptoms, and produce feature importance scores that enhance interpretability. These qualities make it a robust, scalable, and clinically relevant choice for disease prediction tasks.

Hyper parameter Settings:

The Random Forest model set with 100 estimators (decision trees) and a fixed random state of 42 to ensure reproducibility. Other parameters, such as maximum Tree depth and split criteria, were kept at their default settings due to the relatively small dataset size. Advanced hyper parameter optimization methods such as Grid Search or Random Search were postponed for later work, where larger datasets could justify fine-tuning to achieve higher performance.

Evaluation Measures:

The model's performance was evaluated using number of metrics to ensure reliability and comprehensive assessment. These included overall accuracy, which measured the proportion of correctly classified samples, and the confusion matrix, Which supplied thorough breakdown of true positives, true negatives, and false positives and false negatives, calculated for each class? Furthermore, across various disease categories, precision, recall, and F1- score were both strengths and weaknesses across different disease categories. In situations where the test set contained only one class, the metrics were marked as degenerate, acknowledging the limitations caused by insufficient class diversity.

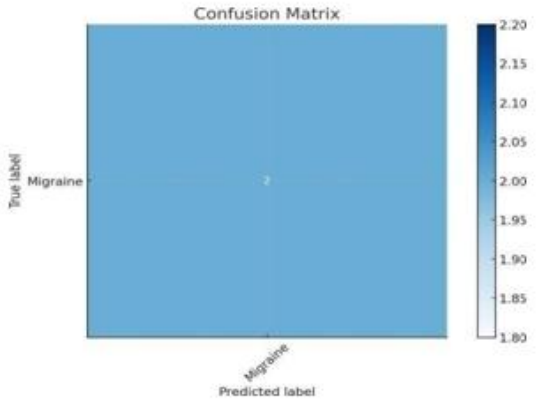

Fig1: Evaluation Matrix

Deployment:

For real-world applicability, the trained Random Forest model (model.pkl) and corresponding class list (classes.pkl) were deployed as a Flask-based web application. The application offers an intuitive interface where users, such as patients or healthcare professionals, can input symptom data in binary format (0 or 1). Built-in input validation mechanisms ensure data correctness prior to prediction. Upon submission, the system outputs the most probable disease along with its confidence level, providing real-time decision support for healthcare applications.

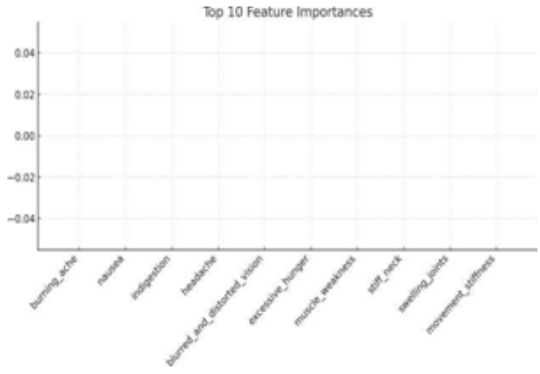

Fig2: Feature Importance Symptoms

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, November 2025
Impact Factor: 7.67

IV. IMPLEMENTATION

The implementation of the proposed disease prediction system focuses on developing an intelligent and interactive platform capable of predicting diseases based on user- selected symptoms and suggesting appropriate remedies. The system was designed utilizing Python, machine learning algorithms, and a Flask-based web interface to ensure both accessibility and efficiency. The dataset that was used for creating a supported model of 31 symptom features associated with specific diseases. Pre- processing steps such as handling missing values and encoding categorical symptom indicators into binary format (where "Yes" was represented as 1 and "No" as 0) were performed to enhance model compatibility. The processed data was divided into training and testing sets to evaluate the model's performance and ensure that its predictions remained consistent and dependable across different samples.

For the predictive modeling, the Random Forest algorithm was chosen due to its resilience and capacity to handle high-dimensional healthcare data effectively. While it was being trained phase, symptom patterns were mapped to corresponding disease classes, and hyper parameter adjustments were made to enhance the model performance. The model performance was viewed using common metrics like accuracy, was evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score, ensuring a balanced assessment of its classification capability. Once the model achieved satisfactory performance, it was serialized using Job lib to allow seamless integration into the web application without requiring retraining.

The finalized system was implemented as a lightweight, web-based application using the Flask framework to enhance usability. Users can access the application directly through a web browser, where they are presented with a list of symptoms in an easy Yes or No format. Upon submission, A portion of responses are transformed into a binary feature vector and processed by the trained Random Forest model to predict the most probable disease. The system then retrieves and displays relevant remedies from a structured knowledge base stored in a CSV file or database. These remedies include simple home treatments and Ayurvedic options, ensuring users receive both diagnostic insights and practical care recommendations. The system was thoroughly tested using multiple symptom combinations to verify prediction accuracy, and user-level testing confirmed that the interface is intuitive, responsive, and effective in providing accurate disease predictions and appropriate remedies in real time.

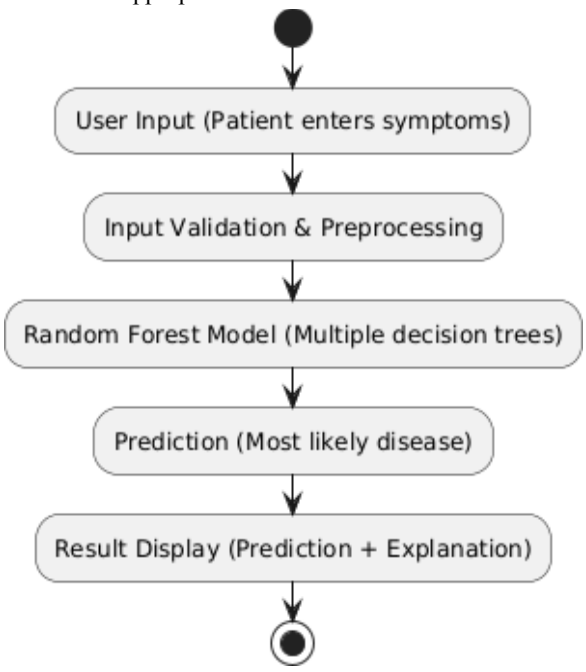

Fig3: Work Flow of System

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, November 2025
Impact Factor: 7.67

V. FUTURE WORK

The proposed disease prediction system has shown promising results; however, there is significant scope for future enhancement. Further development can focus on improving accuracy, interpretability, and real-world applicability of the model. The following points outline key directions for future work:

- Include more diverse datasets from multiple hospitals and regions to enhance prediction reliability.
- Apply techniques like oversampling or synthetic data generation to address class imbalance.
- Integrate explainable AI tools such as SHAP or LIME for better interpretability of predictions.
- Incorporate multimodal data such as medicalimages and lab reports for more comprehensive disease prediction.
- Integrate real-time symptom tracking and feedback mechanisms for continuous model improvement.

VI. CONCLUSION

This project focuses on creating a machine learning system that can predict possible diseases based on simple Yes/No answers to symptoms. Using the Random Forest algorithm, the system can handle overlapping or related symptoms and provide accurate predictions. It was developed through a process of preparing data, training the model, and deploying it on an easy-to-use web interface built with Flask. The tool not only predicts potential diseases but also suggests remedies and shows a confidence score, helping users understand when to seek professional medical advice. Designed to be simple and accessible, it is especially useful for people without medical training or in areas with Limited healthcare access, demonstrating how machine learning can be applied to practical healthcare support.

REFERENCES

[1] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[2] I. Kononenko, “Machine learning for medical diagnosis: history, state of the art and perspective,” Artificial Intelligence in Medicine, vol. 23, no. 1, pp. 89–109, 2001.
[3] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, pp. 115–118, 2017.
[4] A. Rajkomar et al., “Scalable and accurate deep learning with electronic health records,” npj Digital Medicine, vol. 1, no. 18, pp. 1–10, 2018.
[5] E. Choi et al., “Doctor AI: Predicting clinical events via recurrent neural networks,” in Proceedings of Machine Learning for Healthcare (ML4H), 2016.
[6] B. Shickel et al., “Deep learning in healthcare: A review,” IEEE Transactions on Neural Networks and Learning Systems (TNNLS), vol. 29, no. 10, pp. 1–19, 2017.
[7] L.Deyetal.,“Machine learning techniques for medical Diagnosis and prediction of diabetes,”IEEEAccess,vol.6,pp.41374– 41385,2018.
[8] J. Wang et al., “Ensemble methods for disease prediction,” Journal of Healthcare Informatics Research (JHI), vol. 4, pp. 283–295, 2020.
[9] A. Mohammed et al., “Liver disease prediction using hybrid machine learning models,” Procedia Computer Science, vol. 82, pp. 195–202, 2016.
[10] K. Patel etal., “Breast cancer prediction using random forest classifier,” International Journal of Computer Applications (IJCA), vol. 123, no. 16, pp. 32–39, 2015.
[11] E. Tjoa and C. Guan, “A survey on explainable artificial intelligence (XAI): Toward medical XAI,” Information Fusion, vol. 71, pp. 20–45, 2021.
[12] M. Ghassemi, L. Oakden-Rayner, and A. Beam, “The false hope of current approaches to explainable artificial intelligence in health care,” Patterns, vol. 2, no. 9, 100296, 2021.
[13] A.Rajkomar, E.Oren, K.Chen,etal., “Ensuring fairness in machine learning for health,” npj Digital Medicine,vol.4, Art. 302, 2021.
[14] M. T. Nguyen, M. De Coster, and B. De Moor, “Efficient imputation of missing values in medical datasets using ensemble tree-based models,” Applied Soft Computing, vol. 85, 105816, 2020.